在线学习资料支持

您可以在华为企业业务网站获得E-Learning课程、培训教材、产品资料、软件工具、技术案例等：

1、 E-Learning课程：登录*华为在线学习网站*，进入“*华为培训/在线学习*”栏目

免费E-Learning课：对网站所有用户免费开放

职业认证E-Learning课：通过任何一项职业认证即可学习所有职业认证培训E-Learning课程

渠道赋能E-Learning课：对华为企业业务合作伙伴免费开放

2、培训教材：登录*华为在线学习网站*，进入“*华为培训/面授培训*”，在具体课程页面即可下载教材。

华为职业认证培训教材、华为产品技术培训教材。无需注册即可下载

3、华为在线公开课(LVC)：http://support.huawei.com/ecommunity/bbs/10154479.html

企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师公开授课

4、产品资料下载：http://support.huawei.com/enterprise/#tabname=productsupport

5、软件工具下载：http://support.huawei.com/enterprise/#tabname=softwaredownload

更多内容请访问：

- http://learning.huawei.com/cn
- http://support.huawei.com/enterprise/
- http://support.huawei.com/ecommunity/



华为数通认证系列教程-HCDP IERN

部署企业级路由网络

Implementing Enterprise Routing Networks



版权声明

华为数通认证系列教程-HCDP-Enterprise

华为认证数据通信资深工程师-企业级

1.6 版本

华为认证体系介绍

依托华为公司雄厚的技术实力和专业的培训体系，华为认证考虑到不同客户对ICT技术不同层次的需求，致力于为客户提供实战性、专业化的技术认证。

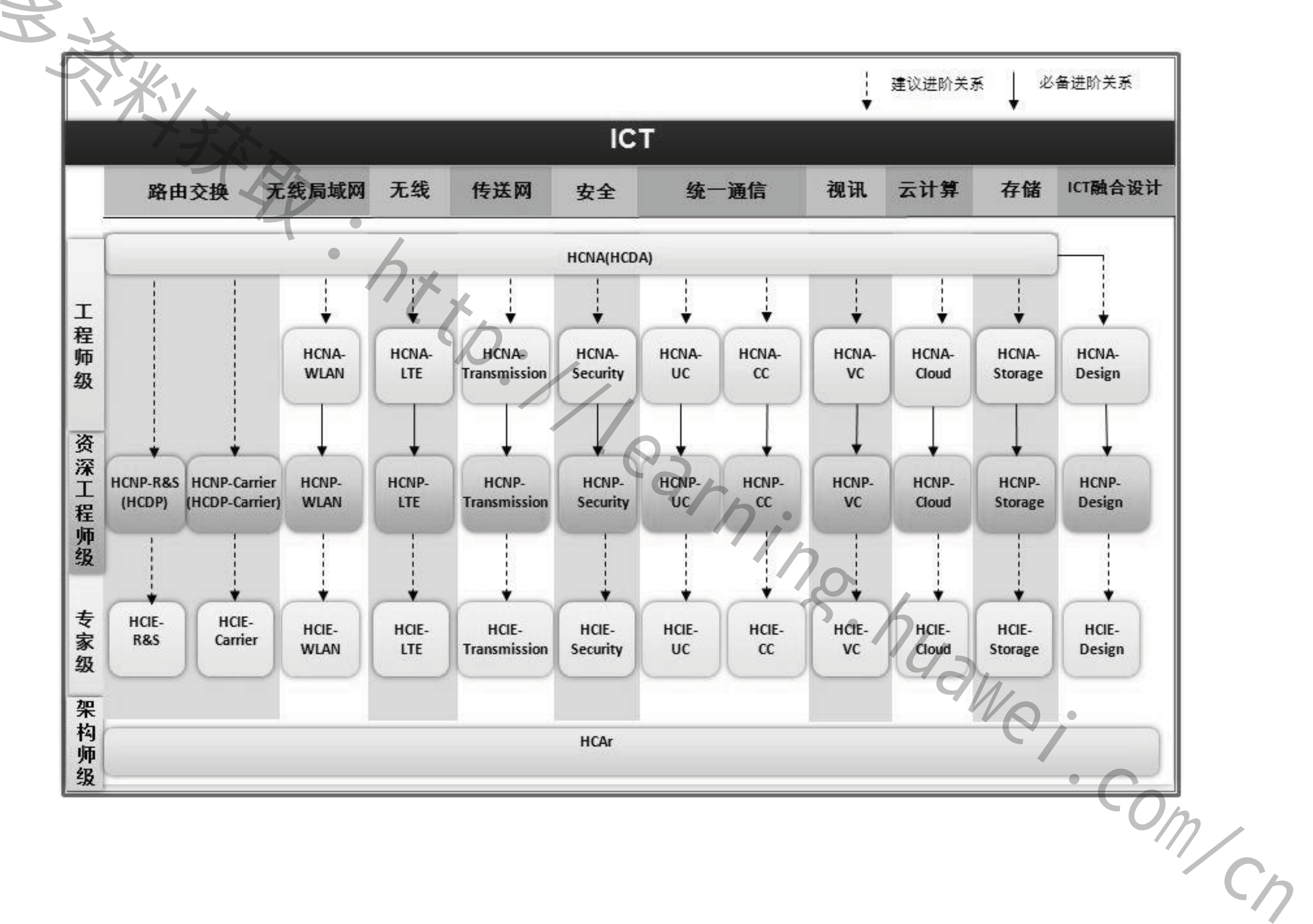
根据ICT技术的特点和客户不同层次的需求，华为认证为客户提供面向十三个方向的四级认证体系。

HCNA(HCDA)认证定位于中小型网络的基本配置和维护。HCNA(HCDA)认证包括但不限于：网络基础知识；流行网络的基本连接方法；基本的网络建造；基本的网络故障排除；华为路由交换设备的安装和调试。通过HCNA(HCDA)认证，将证明您对中小型网络有初步的了解，了解面向中小型企业的网络通用技术，并具备协助设计中小企业网络以及使用华为路由交换设备实施设计的能力。拥有通过HCNA(HCDA)认证的工程师，意味着中小企业有能力完成基本网络搭建，并将基本的语音、无线、云、安全和存储集成到网络之中，满足各种应用对网络的使用需求。

HCNP-Enterprise (HCDP-Enterprise)认证定位于中小型网络的构建和管理。HCNP-Enterprise (HCDP-Enterprise)认证包括但不限于：网络基础知识；交换机和路由器原理；TCP/IP 协议簇；路由协议；访问控制；网络故障的排除；华为路由交换设备的安装和调试。通过 HCNP-Enterprise (HCDP-Enterprise)认证，将证明您对中小型网络有全面深入的了解，掌握面向中小型企业的网络通用技术，并具备独立设计中小企业网络以及使用华为路由交换设备实施设计的能力。拥有通过 HCNP-Enterprise (HCDP-Enterprise)认证的工程师，意味着中小企业有能力完成完整网络的搭建，将企业中所需的语音、无线、云、安全和存储全面地集成到网络之中，并且能满足各种应用对网络的使用需求，进而提供较高的安全性、可用性和可靠性。

HCIE-Enterprise 认证定位于大中型复杂网络的构建、优化和管理。HCIE-Enterprise 认证包括但不限于：不同网络和各种路由器交换机之间的互联；复杂连接问题的解决；使用技术解决方案提高带宽、缩短相应时间、最大限度地提高性能、加强安全性和支持全球应用；复杂网络的故障排除。通过HCIE-Enterprise 认证，将证明您对大型网络有全面深入的了解，掌握面向大型企业网络的技术，并具备独立设计各种企业网络以及使用华为路由交换设备实施设计的能力。拥有通过 HCIE-Enterprise 认证的工程师，意味着大中企业有能力独立完成完整的网络搭建，将企业中所需的语音、无线、云、安全和存储全面地集成到网络之中，并且能满足各种应用对网络的使用需求；能够提供完整的故障排除能力；能根据企业和网络技术的发展，规划企业网络的发展，并提供高安全性、可用性和可靠性。

华为认证协助您打开行业之窗，开启改变之门，屹立在ICT世界的潮头浪尖！



建议进阶关系
必备进阶关系
ICT
路由交换
无线局域网
无线
传送网
安全
统一通信
视讯
云计算
存储
ICT融合设计
工程师级
资深工程师级
专家级
架构师级
HCNA(HCDA)
HCNA-WLAN
HCNA-LTE
HCNA-Transmission
HCNA-Security
HCNA-UC
HCNA-CC
HCNA-VC
HCNA-Cloud
HCNA-Storage
HCNA-Design
HCNP-R&S (HCDP)
HCNP-Carrier (HCDP-Carrier)
HCNP-WLAN
HCNP-LTE
HCNP-Transmission
HCNP-Security
HCNP-UC
HCNP-CC
HCNP-VC
HCNP-Cloud
HCNP-Storage
HCNP-Design
HCIE-R&S
HCIE-Carrier
HCIE-WLAN
HCIE-LTE
HCIE-Transmission
HCIE-Security
HCIE-UC
HCIE-CC
HCIE-VC
HCIE-Cloud
HCIE-Storage
HCIE-Design
HCAr

前言

简介

本书为 HCDP-IERN 认证培训教程，适用于准备参加 HCDP-IERN 考试的学员或者希望系统掌握通用路由协议原理以及在华为通用路由平台 VRP 上的实现的读者。

内容描述

本书共包含五个 Module，由浅入深地介绍了通用路由协议原理、在 VRP 上的配置与实现以及华为路由产品的特点和应用。

Module 1 首先简明扼要地介绍了 IPv4 地址规划和子网划分，帮助读者巩固基础知识；

Module 2、3 分别系统而详尽地介绍内部网关协议 OSPF 以及外部网关协议 BGP 的工作原理以及在 VRP 上的配置和实现。帮助读者全面深入地掌握 IPv4 路由协议知识；

Module 4 通过丰富的实例分析阐述了如何灵活使用各种工具实现路由的控制和选择。帮助读者提高综合规划和灵活使用路由协议的能力。

Module 5 简要介绍了组播地址、IGMP、PIM-DM、PIM-SM 等，帮助读者了解组播基础知识、常用组播协议基本原理以及组播应用。

本书引导读者循序渐进地掌握路由技术在华为产品中的实现，读者也可以根据自身情况选择感兴趣的章节阅读。

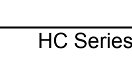
读者必备知识背景

为了更好地掌握本书内容，阅读本书的读者应首先具备以下基本条件之一：

(1)　参加过 HCDA 培训

(2)　通过 HCDA 考试

(3)　熟悉 TCP/IP 协议栈工作原理，熟悉 IP 地址



本书常用图标



IPv6路由器



SOHO路由器



语音模块的路由器


中低端路由器


高端路由器


核心路由器

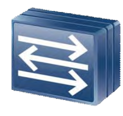

集线器


插座式交换机


汇聚交换机


核心交换机


边缘交换机


堆叠交换机


AP


AP大功率


无线网桥


无线网卡


接入服务器


语音网关

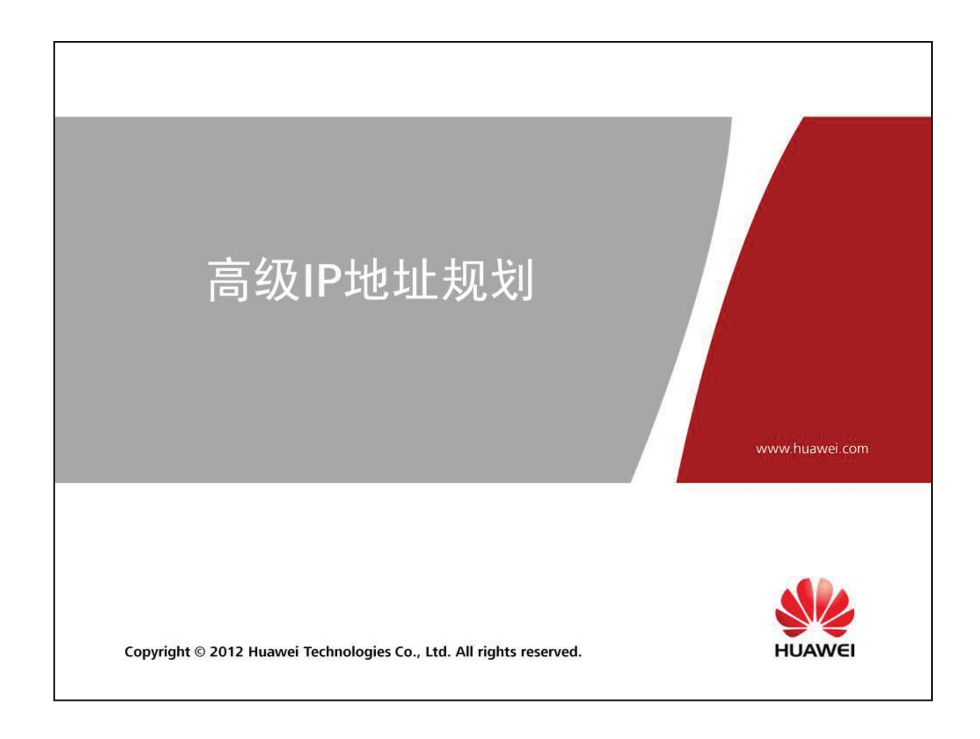

防火墙


网络电话系统

目 录

Module 1

Advanced IP


高级IP地址规划
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

IP地址的合理规划是网络设计中的重要一环，IP地址规划的好坏，直接影响网络路由协议算法的效率和路由收敛的快慢，直接关系网络的稳定性、可扩展性和整体性能，影响到网络的管理。





培训目标

学完本课程后，您应该能：

- 掌握VLSM技术，会用VLSM技术进行子网规划
- 掌握路由聚合与CIDR

Page2





目 录

- IP地址规划的重要性
- 使用VLSM技术进行IP地址规划
- 路由聚合与CIDR





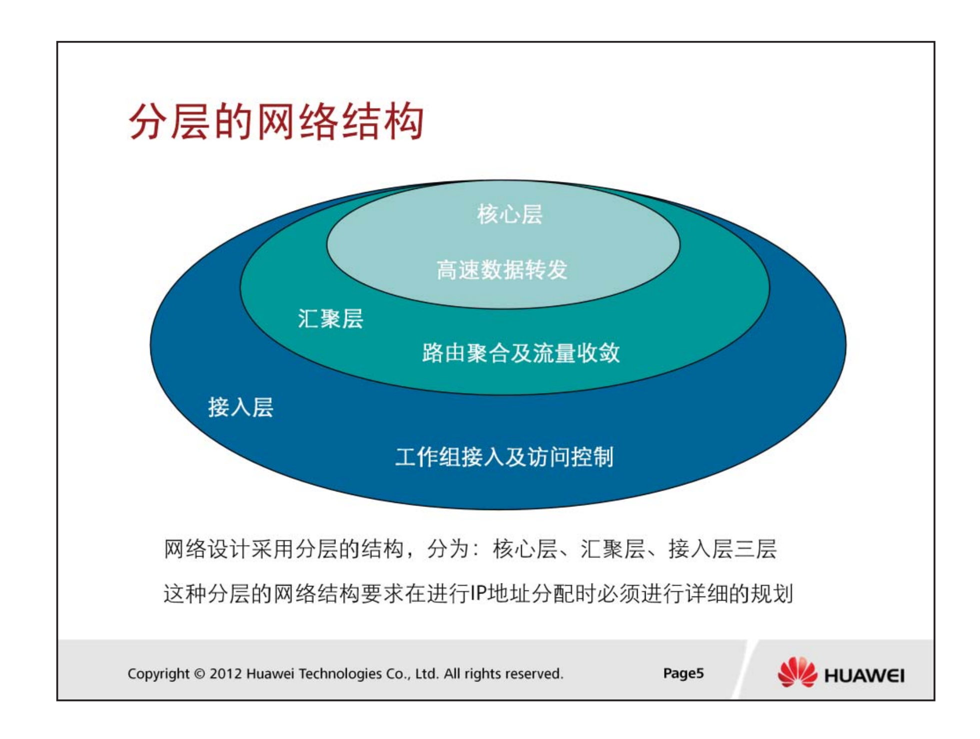
IP地址规划的重要性

使用VLSM技术进行IP地址规划

路由聚合与CIDR

Page4





- 由于网络的规模越来越大，网络出现了一种分层的网络结构，一般的网络可以分为：核心层、汇聚层和接入层，核心层完成数据包的交换，实现高速的数据流量运转，核心层的设备不但需要容量大，转发快，而且需要具备高稳定性；汇聚层的作用是隔离拓朴结构变化、控制路由表的大小及控制网络的收敛，并且实现丰富的业务特性；接入层将终端用户接入到网络中，需要大量的端口，强大的接入能力，实现丰富的业务特性；这种分层的网络结构要求在进行IP地址分配的时候必须进行详细的规划。

IP地址规划的重要性

IP地址规划的好坏，将会直接给网络带来影响：

网络路由协议算法的效率

网络的性能

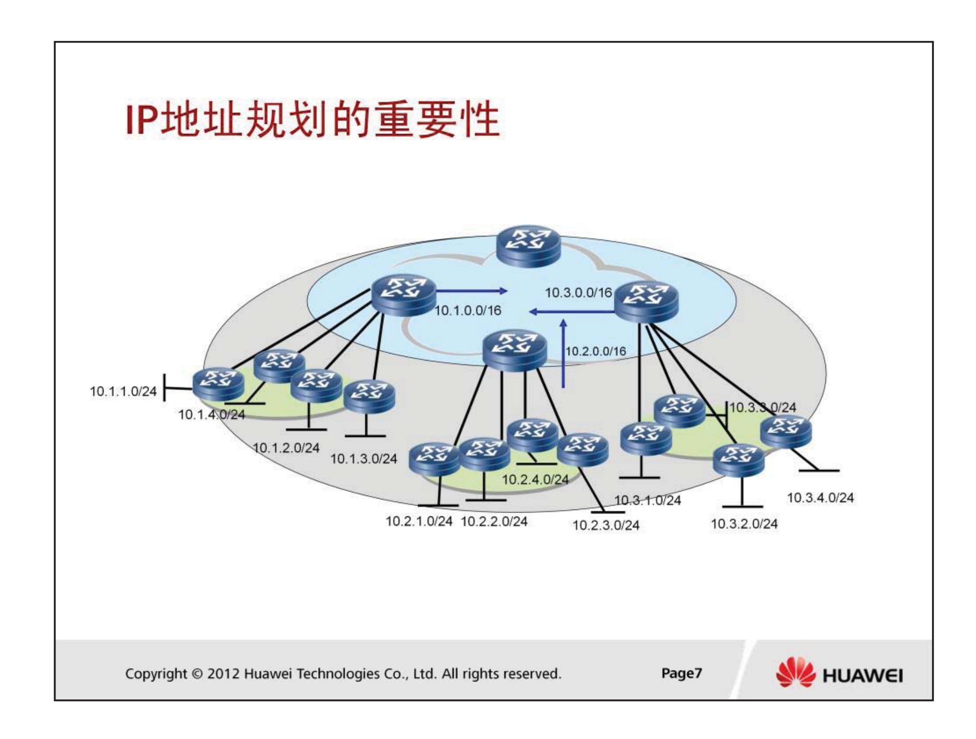
网络的扩展

网络的管理

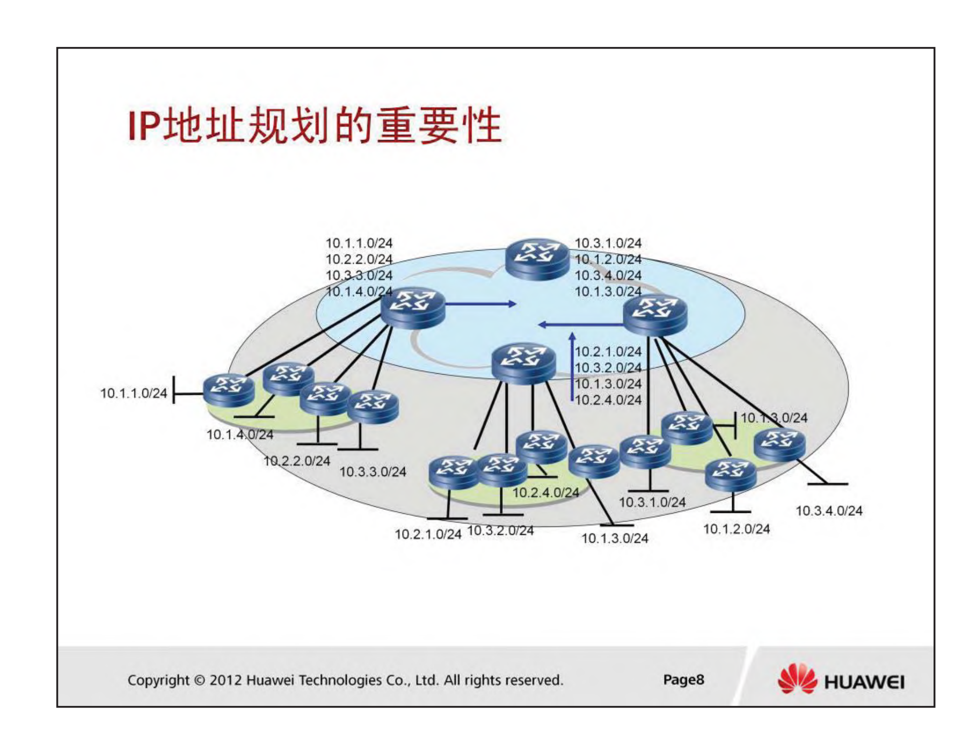
IP地址规划是一项艺术创造！

 Page6 

- IP地址的合理规划是网络设计中的重要一环，大型网络必须对IP地址进行统一规划。
- IP地址规划的好坏，影响到网络路由协议算法的效率，影响到网络的性能，影响到网络的扩展，影响到网络的管理，也必将直接影响到网络应用的进一步发展。
- 通过IP地址规划可以反映出一个网络的规划质量、一个网络设计师的技术水准。



- 图中每个区域有四个网段，在核心层对网段进行聚合，并将聚合后的路由发到其他区域。
- 区域内每台路由器将会有：4条本区域路由和2条其他区域的聚合路由。



- 相对于上一张胶片的组网图，本图因为没有经过详细的IP地址规划，所以在核心路由器上无法聚合，将向其他区域发布本区域中所有的网段路由，每台路由器的路由条目数为12，为上一张胶片中路由条目数的2倍。




Page9



IP地址分类

1.0.0.0~126.255.255.255

0	network(7bits)	host(24bits)	A类地址

128.0.0.0~191.255.255.255

1	0	network(14bits)	host(24bits)	B类地址

192.0.0.0~223.255.255.255

1	1	0	network(21bits)	host(24bits)	C类地址

224.0.0.0~239.255.255.255

1	1	1	0	组播地址	D类地址

240.0.0.0~255.255.255.255

1	1	1	1	0	保留	E类地址

 Page10 HUAWEI

- IP地址长度为32比特，分为网络部分和主机部分。网络部分用于唯一地标识一个物理或者逻辑链路，主机部分用于唯一地标识该链路上的一台设备。
- 那么如何区分IP地址的网络部分和主机部分呢？最初互联网络设计者根据网络规模大小规定了地址类，把IP地址分为A、B、C、D、E五类。
- A类IP地址的网络部分为第一个八位数组（octet），第一个字节的第一位（最左边那一位）为“0”，因此，网络部分的有效位数为7位，这样A类地址的第一个字节为1~126之间（127留作它用）。例如10.1.1.1、126.2.4.78为A类地址。A类地址的主机部分为剩余的三个字节共24位。A类地址的范围为1.0.0.0~126.255.255.255，每一个A类网络共有224个A类IP地址。
- B类IP地址的网络部分为前两个八位数组（octet），第一个字节的第一位为“1”，第二位为”0”。因此，网络部分的有效位数为14位，B类地址的第一个字节为128~191之间。例如128.1.1.1、168.2.4.78为B类地址。B类地址的主机部分为剩余的二个字节16位。B类地址的范围为128.0.0.0~ 191.255.255.255，每一个B类网络共有216个B类IP地址。

- C类IP地址的网络部分为前三个八位数组（octet），第一个字节前二位为“1“，第三位为”0”。因此，网络部分的有效位数为21位，C类地址的第一个字节为192~223之间。例如192.1.1.1、220.2.4.78为C类地址。C类地址的主机部分为剩余的一个字节8位。C类地址的范围为192.0.0.0~ 223.255.255.255，每一个C类网络共有28=256个C类IP地址。
- D类地址第一个字节前三位为“111“，第四位为”0”，因此，D类地址的第一个字节为224~239。D类地址通常作为组播地址。

- E类地址第一个字节为240~255之间，保留用于科学研究。

- 我们经常用到的是A、B、C三类地址。IP地址由国际网络信息中心组织（International Network Information Center，InterNIC）根据公司大小进行分配。过去通常把A类地址保留给政府机构，B类地址分配给中等规模的公司，C类地址分配给小型单位。然而，随着互联网络的飞速发展，再加上IP地址的浪费，导致现在IP地址已经非常紧张。

私有IP地址

私有IP地址

- 10.0.0.0~10.255.255.255
- 172.16.0.0~172.31.255.255
- 192.168.0.0~192.168.255.255

　Page12



- 在进行IP地址规划时，我们通常在公司内部网络使用私有IP地址。私有IP地址是由InterNIC预留给各个企业内部网络自由支配的IP地址。使用私有IP地址不能直接访问Internet。原因很简单，私有IP地址不能在公网上使用，公网上没有针对私有地址的路由，如果在公网上使用将会产生地址冲突问题。当访问Internet时，需要利用网络地址转换（NAT，Network Address Translation）技术，把私有IP地址转换为Internet可识别的公有IP地址。InterNIC预留了以下网段作为私有IP地址：A类私有地址为10.0.0.0~10.255.255.255; B类私有地址为172.16.0.0~172.31.255.255; C类私有地址为192.168.0.0~192.168.255.255等。
- 使用私有IP地址，不仅减少了企业用于购买公有IP地址的投资，而且节省了IP地址资源。但是这并不能完全解决IP地址短缺问题，目前已经正式提出了IPv6协议。在IPv6地址中有128个二进制位，共约2128个IP地址，完全可以解决IP地址紧张问题。

特殊IP地址

网络部分	主机部分	地址类型	用 途
Any	全 “0”	网络地址	代表一个网段
Any	全 “1”	广播地址	特定网段的所有节点
127	any	环回地址	环回测试
全 “0”		所有网络	华为Quidway路由器用于指定默认路由
全 “1”		广播地址	本网段所有节点

 Page13 

- IP地址用于唯一的标识一台网络设备，但并不是每一个IP地址都是可用的，一些特殊的IP地址被用于特殊的用途，不能用于标识网络设备。
- 对于主机部分全为“0”的IP地址，称为网络地址，网络地址用来标识一个网段。例如，A类地址1.0.0.0，私有地址10.0.0.0, 192.168.1.0等。
- 对于主机部分全为“1”的IP地址，称为网段广播地址，广播地址用于标识一个网络的所有主机。例如，10.255.255.255 , 192.168.1.255等，路由器可以在10.0.0.0或者192.168.1.0等网段转发广播包。广播地址用于向本网段的所有节点发送数据包。
- 对于网络部分为127的IP地址，例如127.0.0.1往往用于环回测试。
- 全“0”的IP地址0.0.0.0代表所有的主机，华为Quidway系列路由器使用0.0.0.0地址指定默认路由。
- 全“1”的IP地址255.255.255.255，也是广播地址，但255.255.255.255代表所有主机，用于向网络的所有节点发送数据包。这样的广播数据包不能被路由器转发。
- 如上所述，每一个网段会有一些IP地址不能用作主机IP地址。下面让我们来计算一下可用的IP地址。例如B类网段172.16.0.0, 有16个主机

- 位,因此有216个IP地址，去掉一个网络地址172.16.0.0，一个广播地址172.16.255.255不能用作标识主机，那么共有216-2个可用地址。C类网段192.168.1.0，有8个主机位，共有28个IP地址，去掉一个网络地址192.168.1.0，一个广播地址192.168.1.255，共有254个可用主机地址。现在，我们可以这样计算每一个网段的可用主机地址：假定这个网段的主机部分位数为n，那么可用的主机地址个数为2n-2个。
- 网络层设备（例如路由器等）使用网络地址来代表本网段内的主机，可以大大减少路由器的路由表条目。

子网掩码介绍

使用子网掩码（subnet mask）区分网络部分和主机部分

子网掩码使用与IP地址一样的格式

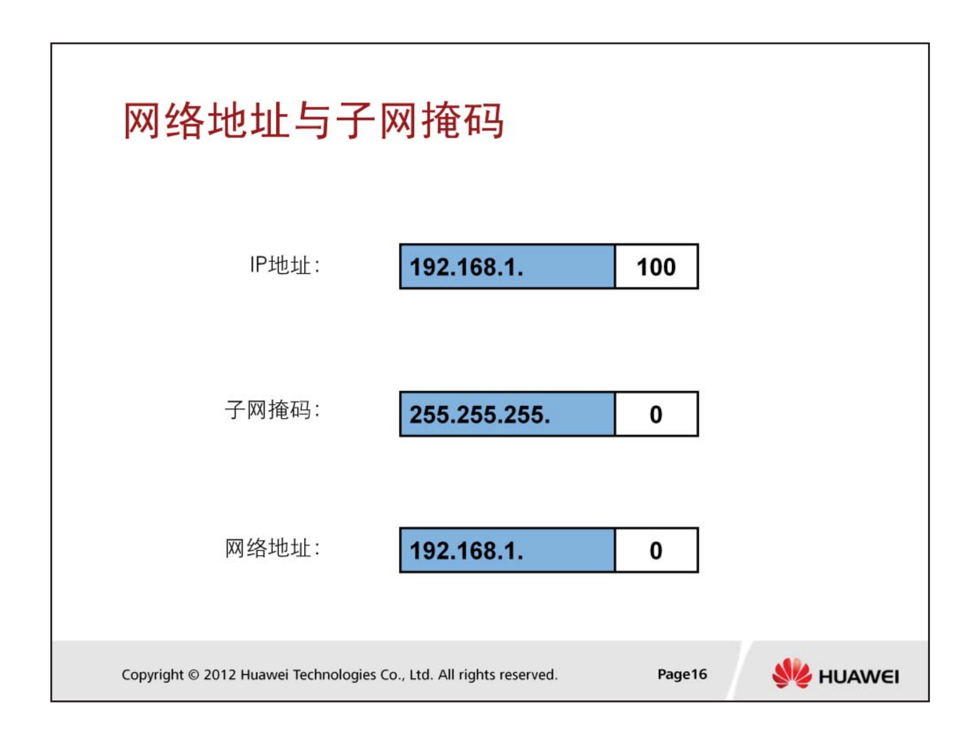
子网掩码的网络部分和子网部分全都是1，主机部分全都是0

例如：B类网络的子网掩码为255.255.0.0

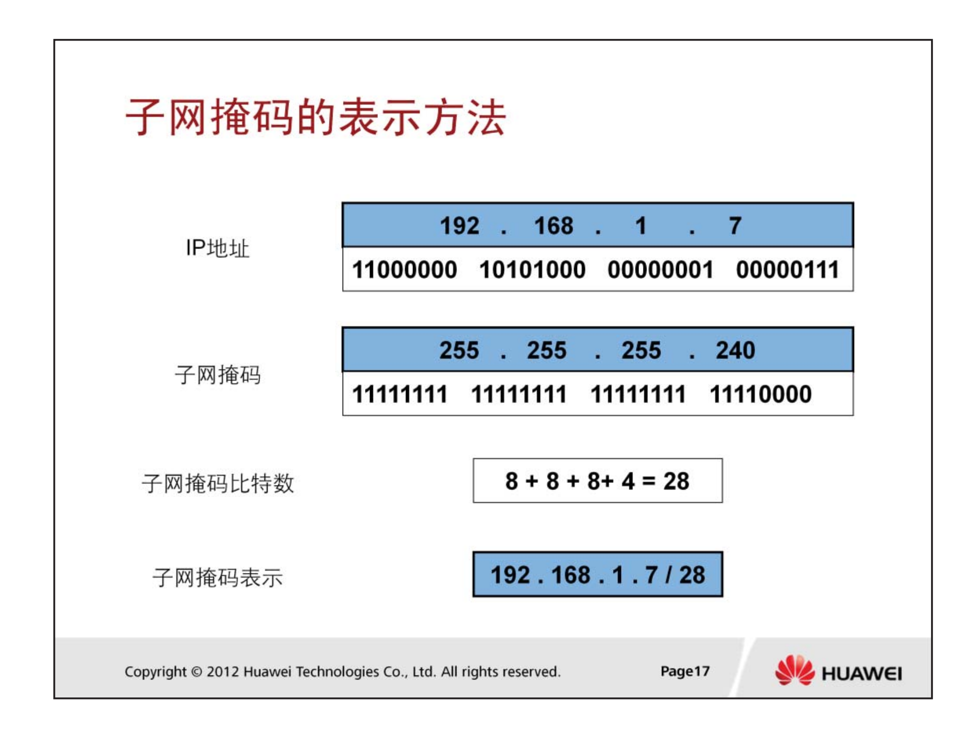
Page15



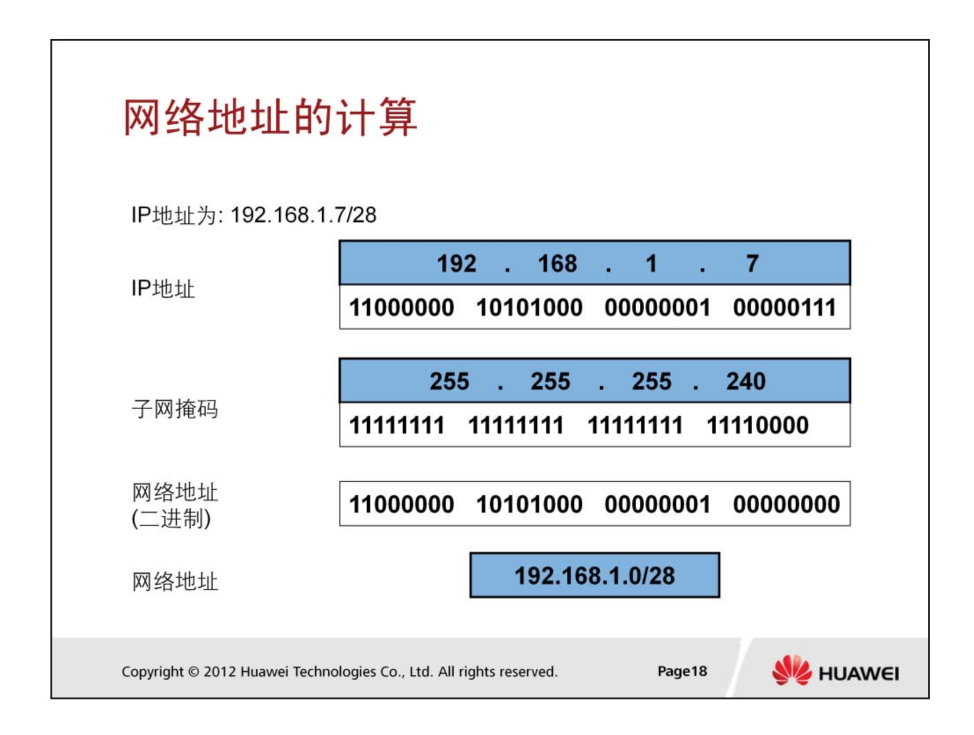
- 缺省状态下，如果没有进行子网划分，A类网络的子网掩码为255.0.0.0，B类网络的子网掩码为255.255.0.0，C类网络子网掩码为255.255.255.0。利用子网掩码进行子网划分，可以使网络地址的使用更有效。对外仍为一个网络，对内部而言，则分为不同的子网。



- 这是一个C类地址，前24bits是网络位，后8bits是主机位。



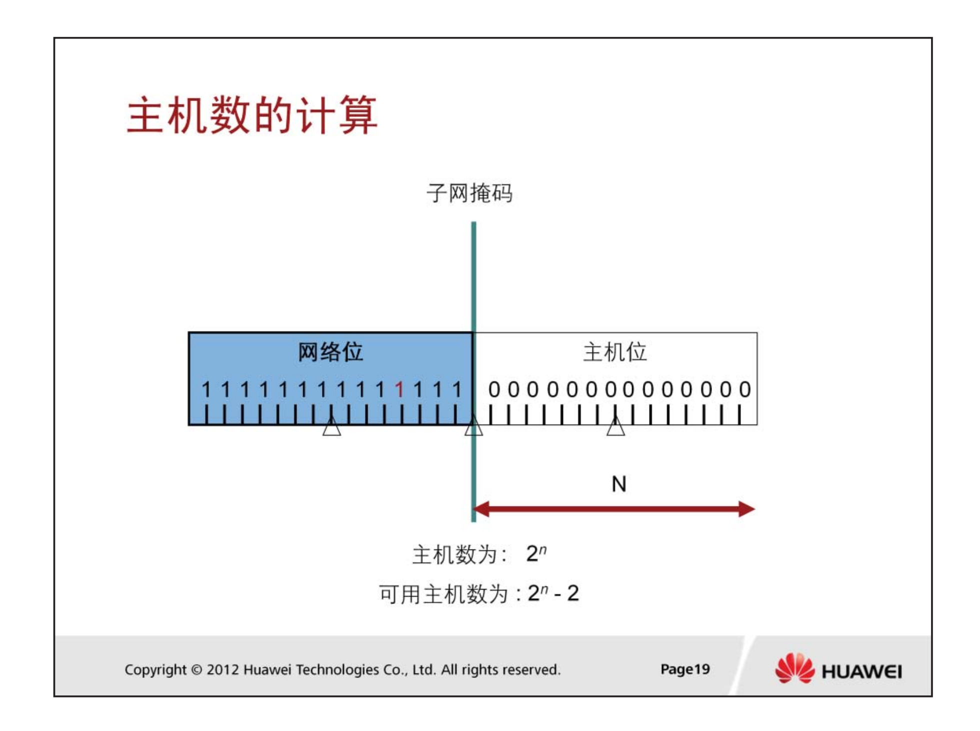
- 如上例所示，子网掩码有两种表示方法。子网掩码255.255.255.240和/28都表示前28bits为网络号。



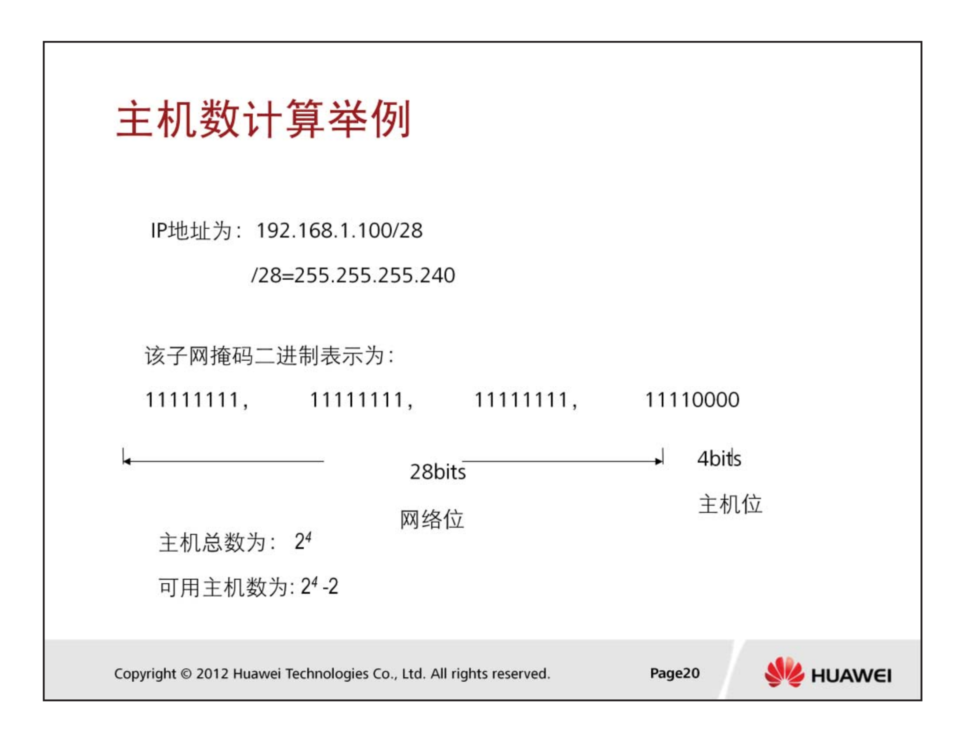
- 如胶片中所示，IP地址和子网掩码都已经知道，那么网络地址就是将IP地址的二进制和子网掩码的二进制进行“与”的计算的结果。“与”的计算方法是1&1=1，1&0=0，0&0=0。那么胶片中IP地址和子网掩码与计算的结果为：

- 11000000，10101000，00000001，00000111
- & 1111111 ，11111111，11111111，11110000
- 11000000，10101000，00000001，00000000

- 最后得到的就是网络地址。



- 主机数的是通过子网掩码来计算的，首先我们要看这个子网掩码中最后有多少位是0。如上图，假设最后有N位为0，那么总的主机数为2n个，可用主机的个数我们要减去全0的网络地址和全1的广播地址，既2n – 2个。



- 本例说明主机数的计算。
- A类地址标准的子网掩码为255.0.0.0，即有24bits的主机位；B类地址的标准子网掩码为255.255.0.0,即有16bits的主机位；C类地址的标准子网掩码为255.255.255.0,即有8bits的主机位。
- C类地址的标准子网掩码有8bits的主机位，但本例中这8bits中的前4bits也用作子网掩码，则所能容纳的主机总数为2的8-4次方，8指的是标准子网掩码的主机位个数，4为用于子网掩码的bits个数，进行相减后，就得到了实际的主机位数，即可表示为28-4，由此可以得到主机总数。

子网数计算举例

IP地址为：192.168.1.100/28

/28=255.255.255.240

该子网掩码二进制表示为：

11111111， 11111111， 11111111， 11110000

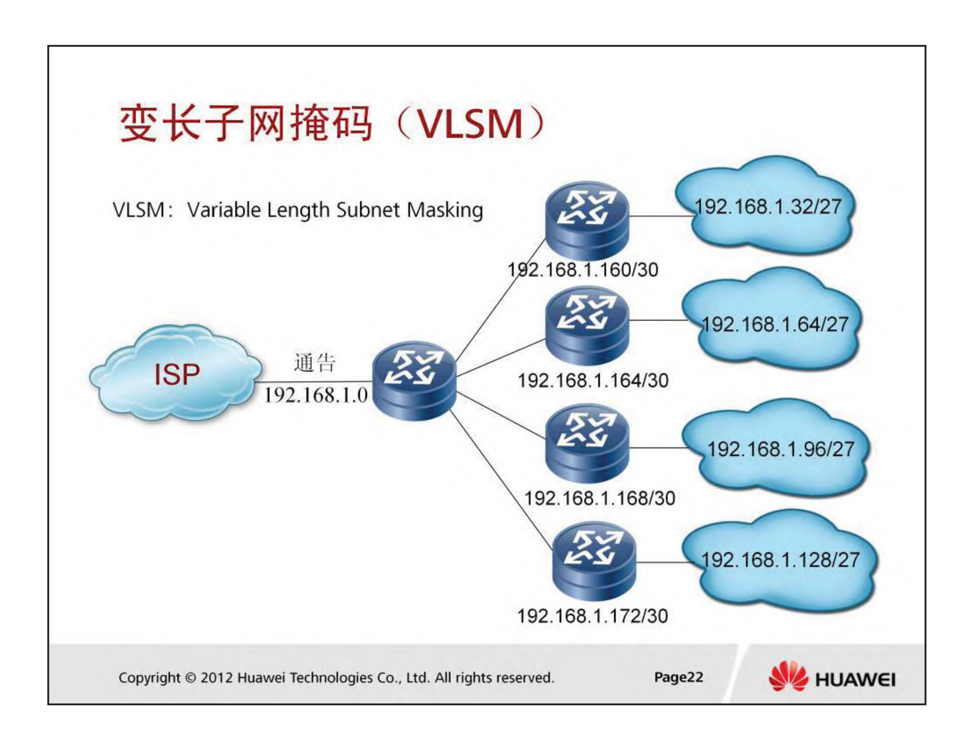
28bits 4bits

网络位 主机位

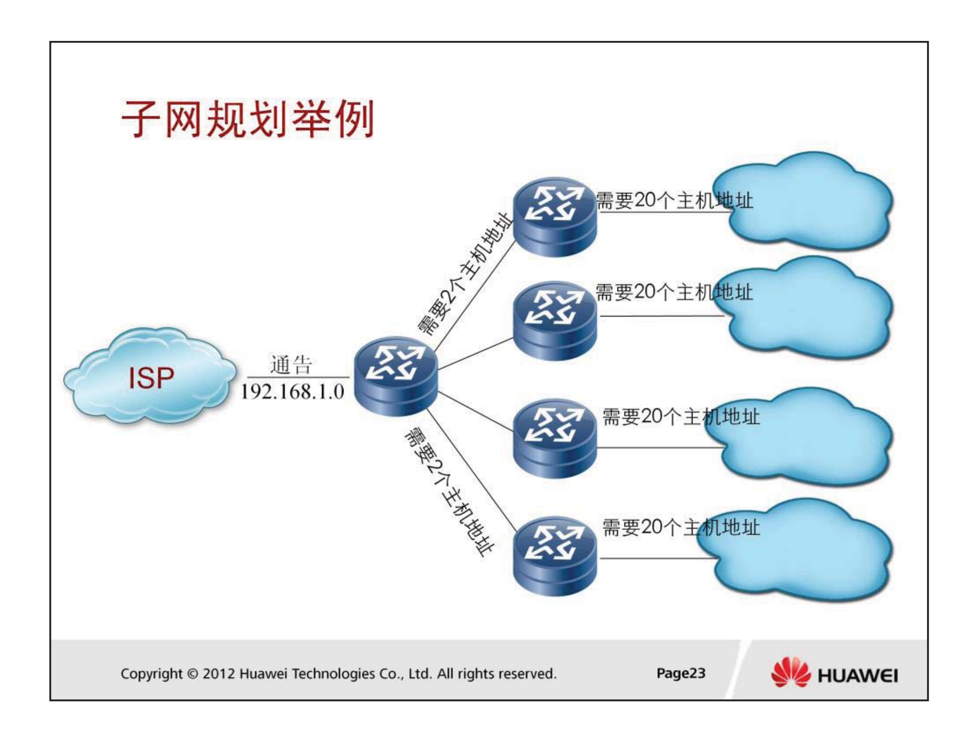
子网总数为：2^{8-4}

Page21 HUAWEI

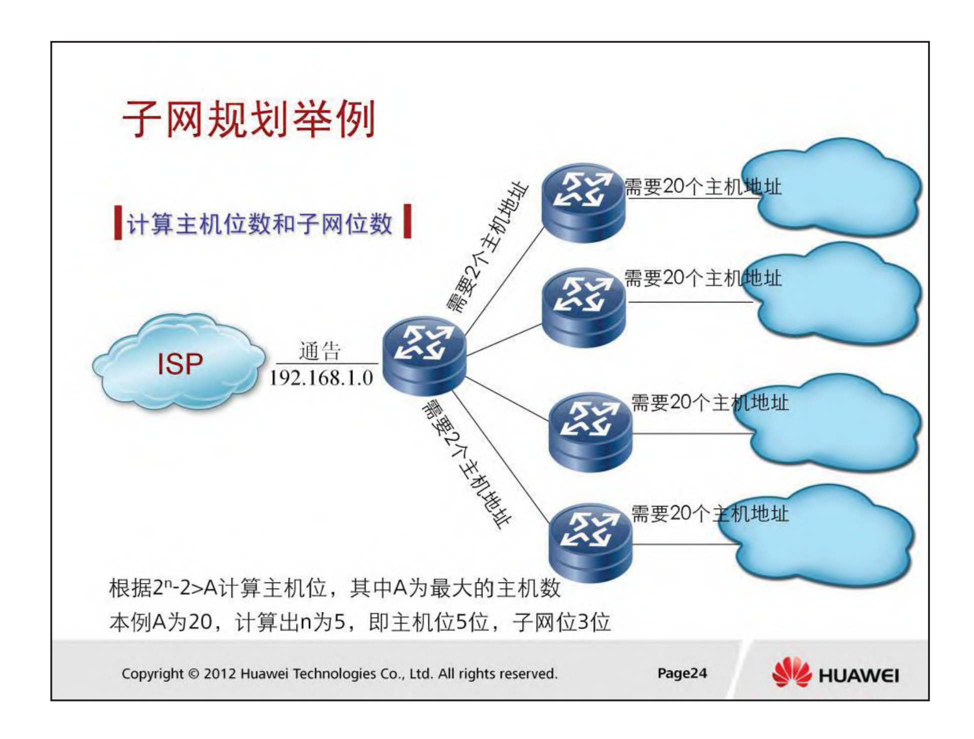
- 我们知道A类地址标准的子网掩码255.0.0.0，也就是说有24bits的主机位；B类地址的标准子网掩码为255.255.0.0，也就是说16bits的主机位；C类地址的标准子网掩码255.255.255.0，也就是说8bits的主机位。
- 胶片中的例子是一个C类地址，标准子网掩码有8bits的主机位，那么计算子网总数的时候就为2的8-4次方，8指的是标准子网掩码的主机位个数，4为实际主机位个数，进行相减后，就得到了子网位数，既可表示为28-4,那么就得到了子网总数。A，B类IP地址以此类推。



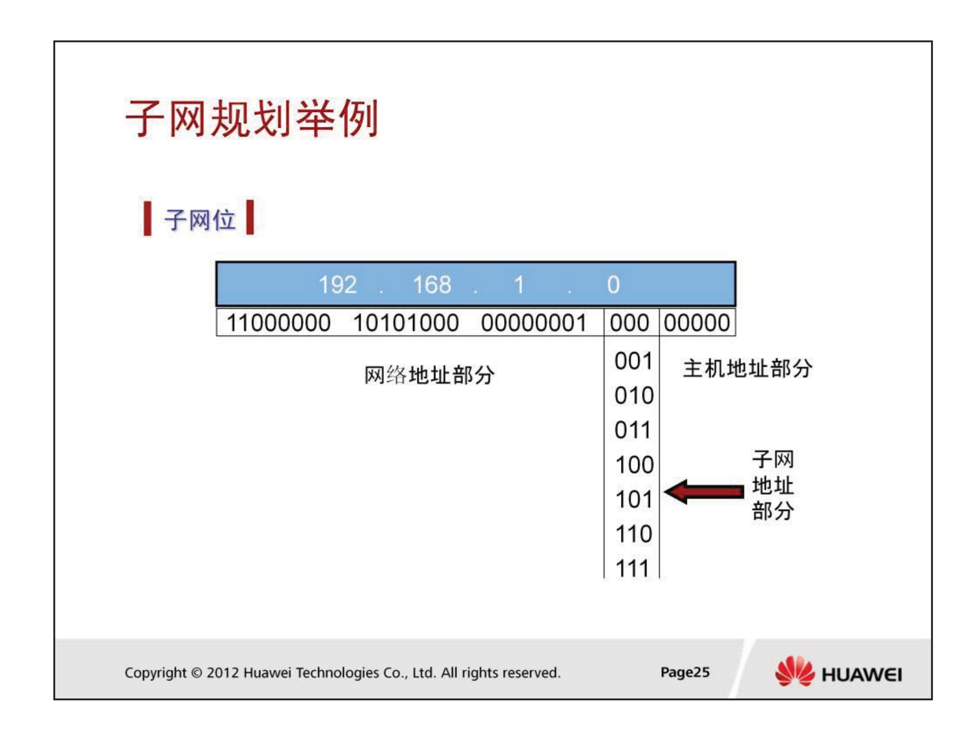
- 把一个网络划分成多个子网，要求每一个子网使用不同的网络标识ID。但是每个子网的主机数不一定相同，而且相差很大，如果我们每个子网都采用固定长度子网掩码，每个子网上分配的地址数相同，这就造成地址的大量浪费。
- 这时候我们可以采用变长子网掩码（VLSM，Variable Length Subnet Masking）技术，对节点数比较多的子网采用较短的子网掩码，子网掩码较短的地址可表示的网络/子网数较少，而子网可分配的地址较多；节点数比较少的子网采用较长的子网掩码，可表示的逻辑网络/子网数较多，而子网上可分配地址较少。这种寻址方案必能节省大量的地址，节省的这些地址可以用于其它子网上。



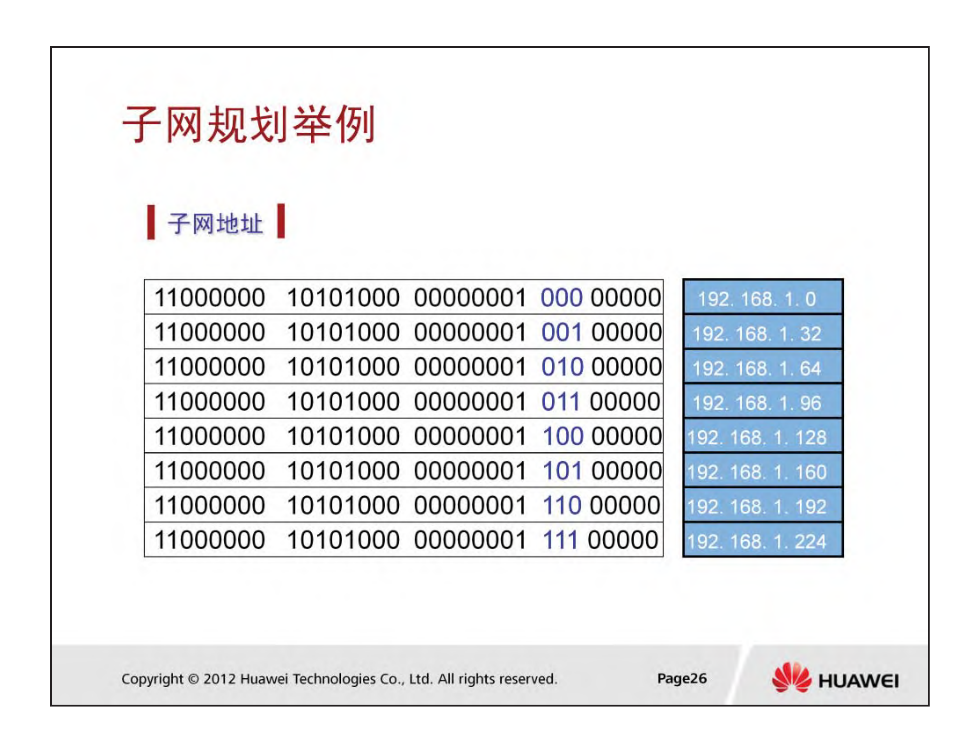
- 如上图所示，某公司准备用C类网络地址192.168.1.0进行IP地址的子网规划。这个公司共购置了5台路由器，一台路由器作为企业网的网关路由器接入当地ISP，其它4台路由器连接四个办公点，每个办公点有20台PC，需要20个主机地址。如何规划IP地址才能满足该公司的组网要求呢?



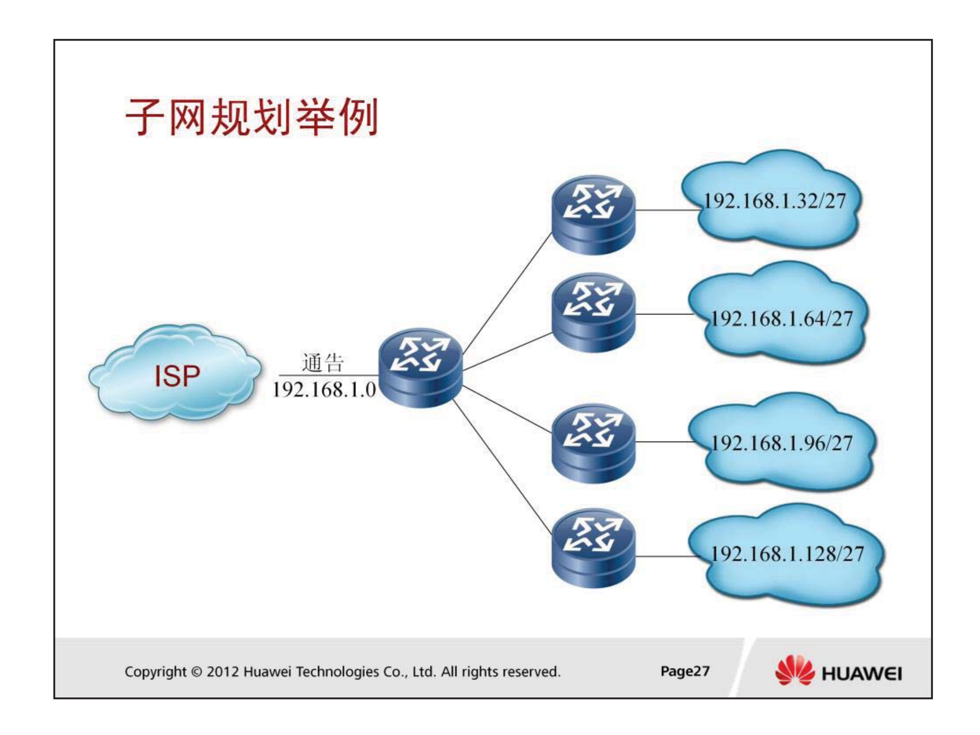
- 子网规划步骤1：确定需要多少个子网，每个子网需要多少个主机，根据公式2n-2>A（A为最大的主机数）计算出子网位和主机位。
- 从上图可以看出，需要划分8个子网，4个办公点网段需要21个IP地址（包括一个路由器接口），与网关路由器相连的4个网段需要2个IP地址。本例先规划出4个办公点的IP地址，然后再规划出4台办公点路由器与网关路由器间的IP地址。
- 根据2n-2>A，本例A为20，计算出n为5，即主机位5位，子网位为3位。因此，4个办公点的主机位为5位。



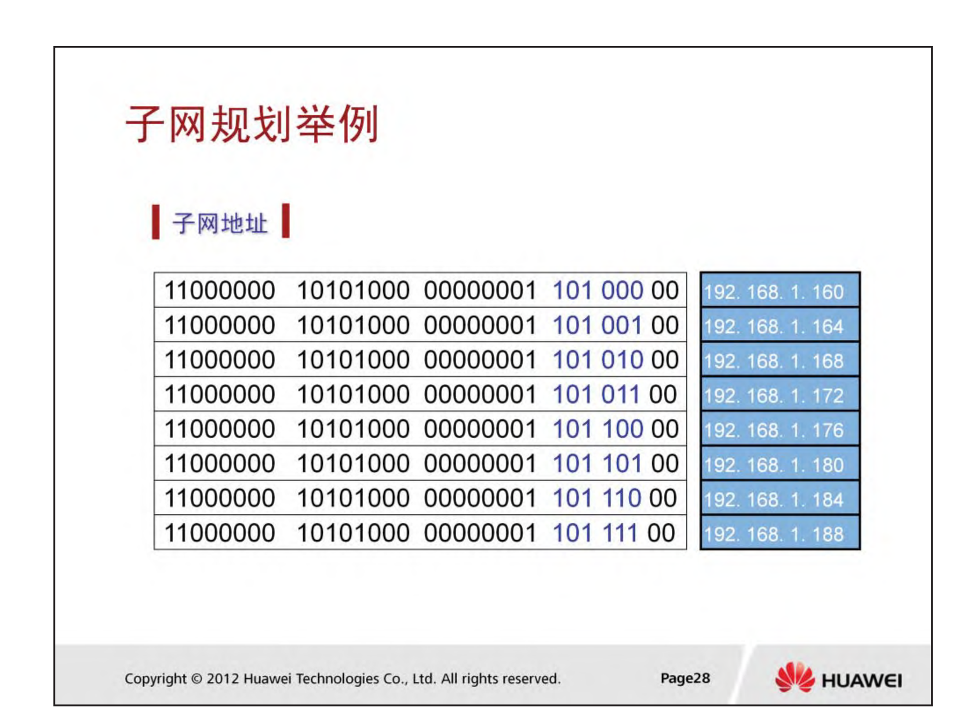
- 步骤2：计算子网位，将192.168.1.0的主机部分分为子网部分和新的主机地址部分，根据步骤1的计算结果，子网位为3位，用二进制表示如上图所示，垂直线标记了子网空间，从二进制000开始计数，将子网位的所有组合列出。



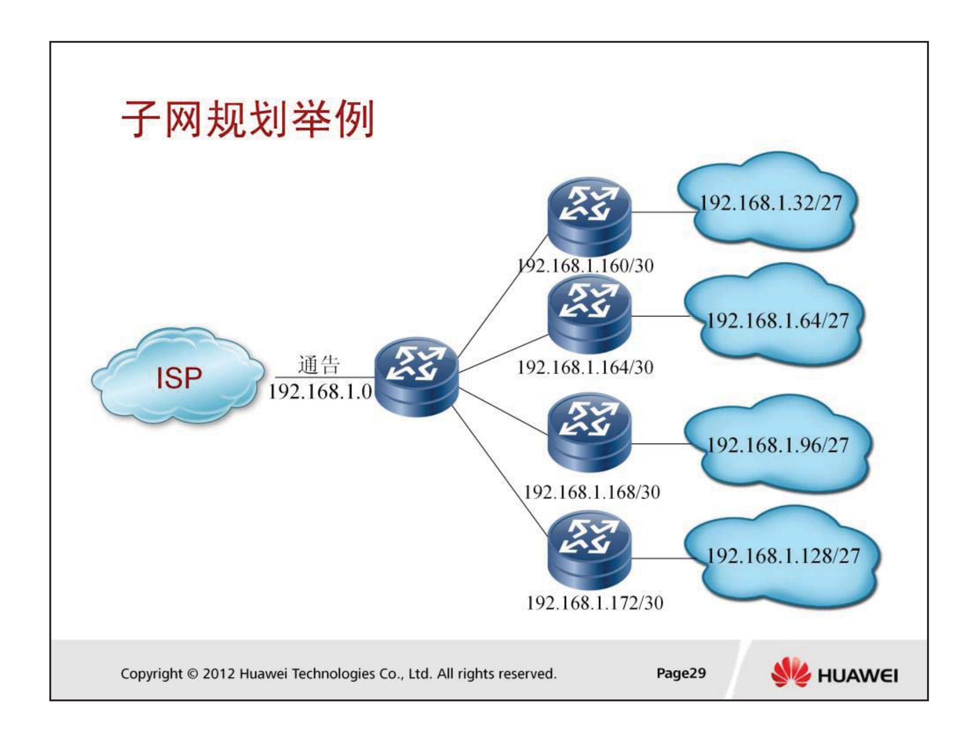
- 步骤3：计算子网地址，将步骤2的结果用点分十进制的格式表示出来，就可以得到图中右边的网段地址。



- 根据步骤3推导出的网段地址，选取其中连续的几个作为最终的结果，本例选取网段192.168.1.32/27，192.168.1.64/27，192.168.1.96/27，192.168.1.128/27，如图中所示。



- 选取网段192.168.1.160规划出新的子网，作为4个办公点路由器和网关路由器之间的子网地址，计算过程同上，可以计算出，4个办公路由器与网关路由器间的子网地址如上图所示。



- 最终的子网规划的结果如图所示。

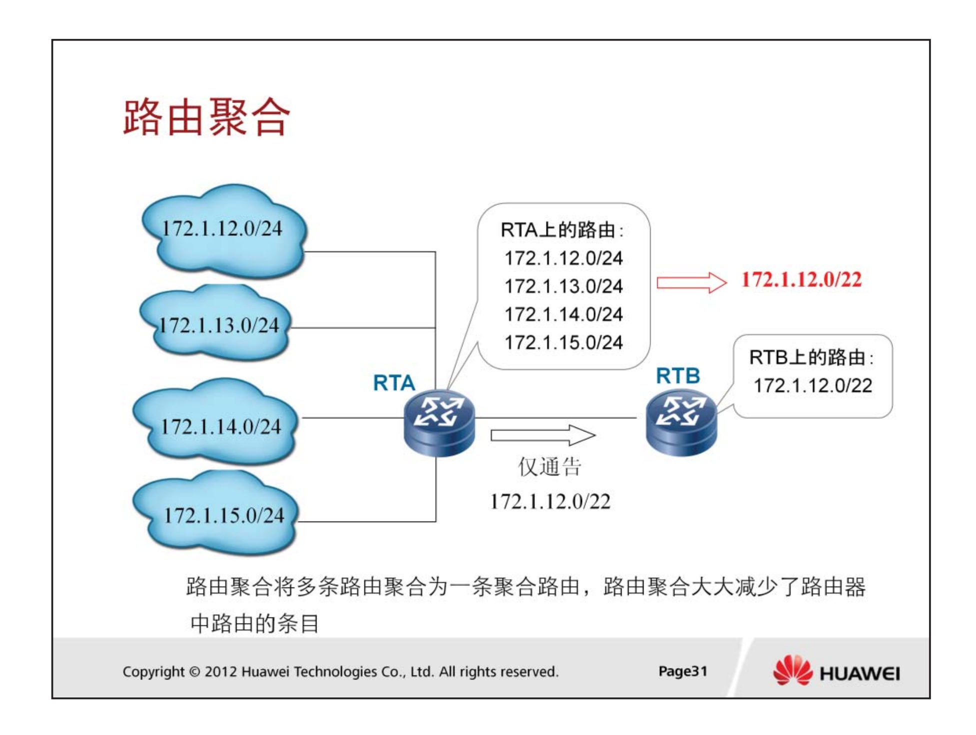


目 录

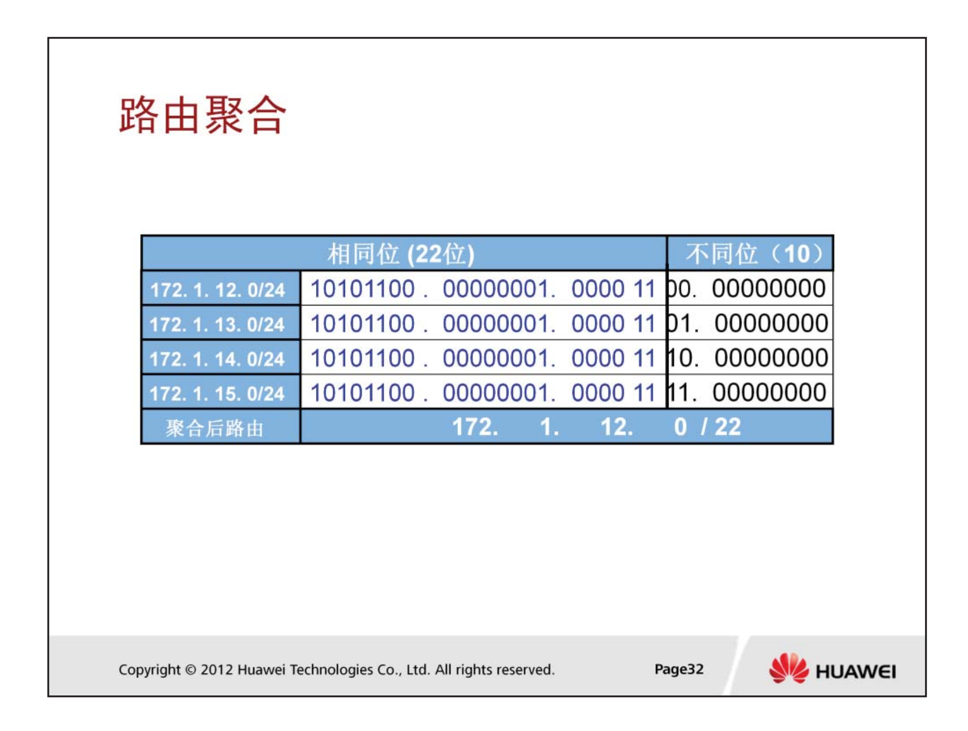
- IP地址规划的重要性
- 使用VLSM技术进行IP地址规划
- **路由聚合与CIDR**

Page30





- 如上图所示，路由器RTA下接4个网段，172.1.12.0/24、172.1.13.0/24、172.1.14.0/24、172.1.15.0/24，那么在路由器RTA上存在这4个网段的路由，在RTA上做路由聚合，可以将这4个网段的路由聚合为一条路由，172.1.12.0/22，在向路由器RTB通告时仅通告172.1.12.0/22这条路由，这样可以大大减少路由的条目数。
- 路由聚合是将多条路由聚合为一条聚合路由，路由聚合可以大大减少路由器中路由的条目数，减轻路由器维护路由条目数的负担，提高网络的利用率。



- 如本例所示，172.1.12.0/24，172.1.13.0/24，172.1.14.0/24，172.1.15.0/24可以聚合为172.1.12.0/22。

无类域间路由（CIDR）

CIDR (Classless Inter Domain Routing)

CIDR使用VLSM技术，突破了传统IP地址分类边界

把路由表中的若干条路由汇聚为一条路由，减少了路由表的规模

支持CIDR的路由协议有：

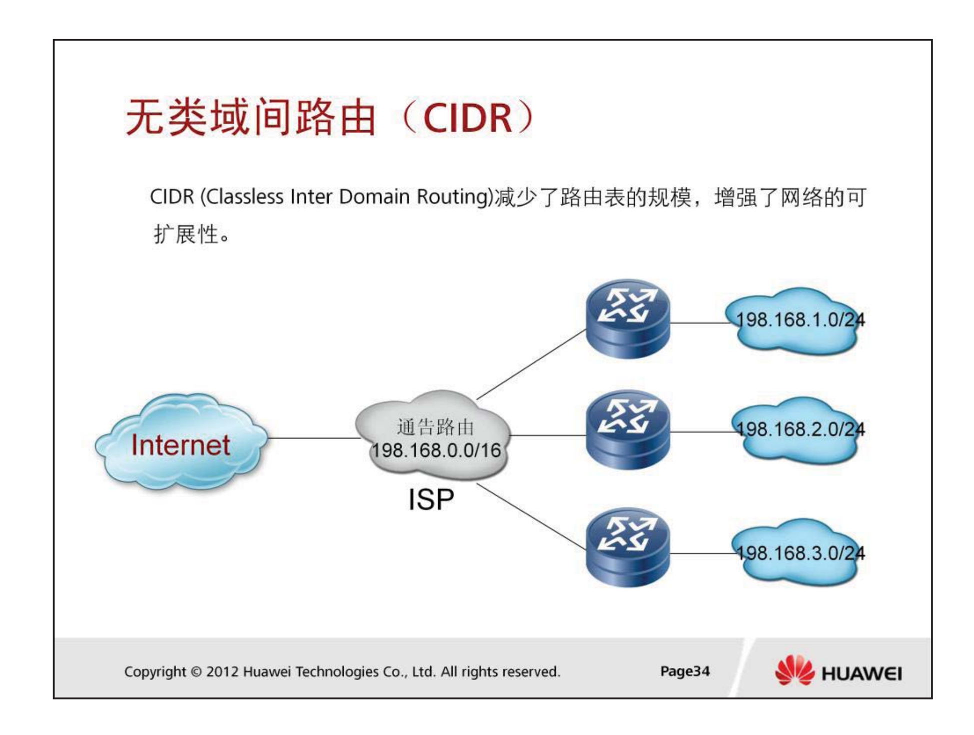
- RIPv2、OSPF、Integrated ISIS、BGPv4

 Page33



- CIDR (Classless Inter Domain Routing)即无类域间路由，由RFC1817定义。CIDR突破了传统IP地址分类边界，使用VLSM技术，把路由表中的若干条路由汇聚为一条路由，减少了路由表的规模，提高了路由器的可扩展性。

- 支持CIDR的路由协议有：RIPv2、OSPF、Integrated ISIS、BGPv4。



- 如上图所示，一个ISP被分配了一些C类网络，198.168.0.0~198.168.255.0。这个ISP准备把这些C类网络分配给各个用户群，目前已经分配了三个C类网段给用户。如果没有实施CIDR技术，ISP的路由器的路由表中会有三条下连网段的路由条目，并且会把它通告给Internet上的路由器。通过实施CIDR技术，我们可以在ISP的路由器上把这三条路由198.168.1.0/24，198.168.2.0/24，198.168.3.0/24汇聚成一条路由198.168.0.0/16。这样ISP路由器只向Internet通告198.168.0.0/16这一条路由，大大减少了路由表的条目数。

- 通常情况下，使用CIDR技术汇聚的网络地址的比特位必须是一致的，如上例所示。如果上图所示的ISP又连接了一个172.178.1.0/24的网段，那么这些网段路由将无法汇聚，无法实现CIDR技术。
- （注：极端情况下，可以汇聚成0.0.0.0/0发布）



问 题

什么是VLSM技术?

用VLSM技术进行子网规划可以分为哪几步?

什么是路由聚合与CIDR?

 Page35



- Q:什么是VLSM技术?
- A:VLSM即Variable Length Subnet Masking，可变长子网掩码，即子网不再采用固定长度的子网掩码，而是可以根据实际情况进行变化，对节点数比较多的子网采用较短的子网掩码，子网掩码较短的地址可表示的网络/子网数较少，而子网可分配的地址较多；节点数比较少的子网采用较长的子网掩码，可表示的逻辑网络/子网数较多，而子网上可分配地址较少。这种寻址方案必能节省大量的地址，节省的这些地址可以用于其它子网上。

- Q: 用VLSM技术进行子网规划可以分为哪几步?
- A: 用VLSM技术进行子网规划大致可以分为4步，第1步，确定子网位数和主机位数；第2步计算子网位；第3步，计算子网地址；第4步，选取子网地址，得出子网规划的结果。

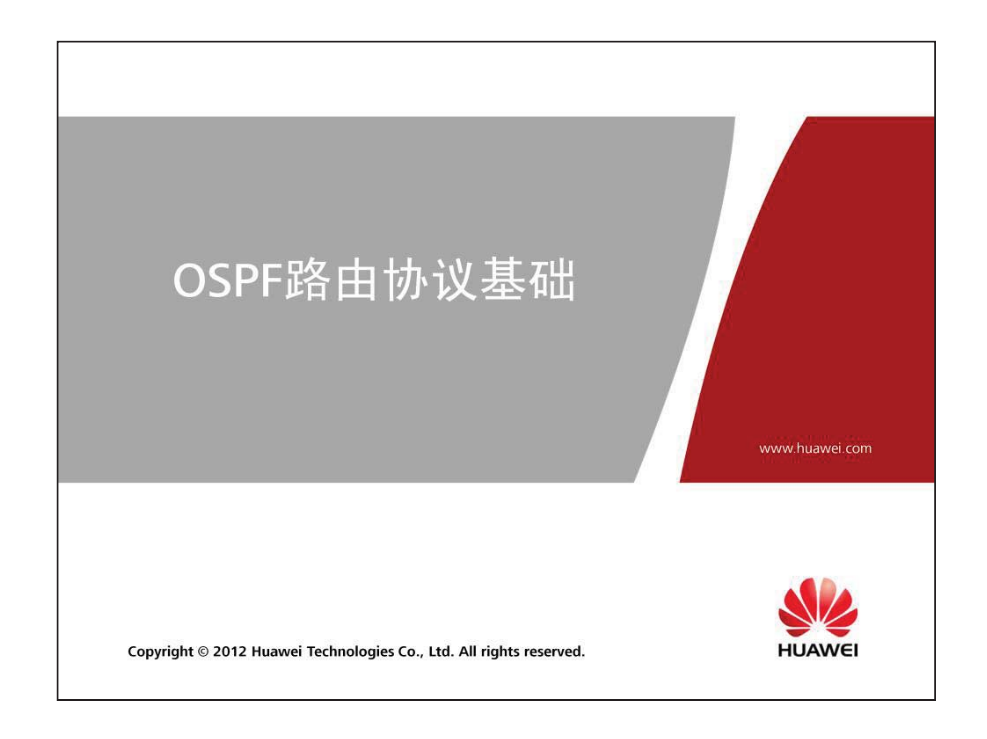
- Q: 什么是路由聚合与CIDR?
- A: 路由聚合指将多条路由聚合为一条聚合路由，路由聚合可以大大减少路由器中路由的条目数，减轻路由器维护路由的负担，提高网络的利用率。

- CIDR即Classless Inter Domain Routing，无类别域间路由，CIDR使用VLSM技术，突破了传统IP地址分类边界，采用CIDR可以把路由表中的若干条路由汇聚为一条路由，减少了路由表的规模。


谢谢
www.huawei.com

Module 2

OSPF


OSPF路由协议基础
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍TCP/IP路由协议之开放式最短路径优先协议（OSPF）的基本概念与基础配置。

OSPF是内部网关协议的一种，基于链路状态算法。

Page1





培训目标

学完本课程后，您应该能：

- 了解OSPF协议基本特点
- 理解链路状态算法的路由计算过程
- 掌握OSPF基本概念
- 掌握OSPF协议的基础配置

 Page2



OSPF基本特点

支持无类域间路由（CIDR）

无路由自环

收敛速度快

使用IP组播收发协议数据

支持多条等值路由

支持协议报文的认证

Page3



OSPF基本特点如下:

支持无类域间路由（CIDR）:

OSPF是专门为TCP/IP环境开发的路由协议，支持无类域间路由（CIDR）和可变长子网掩码（VLSM）。

无路由自环:

由于路由的计算是基于详细链路状态信息（网络拓扑信息）的，所采用的SPF算法本身不会产生环路，并且OSPF报文携带生成者的ID信息，因此OSPF计算的路由无自环。

收敛速度快:

触发式更新，一旦拓扑结构发生变化，新的链路状态信息立刻泛洪，对拓扑变化敏感。

使用IP组播收发协议数据:

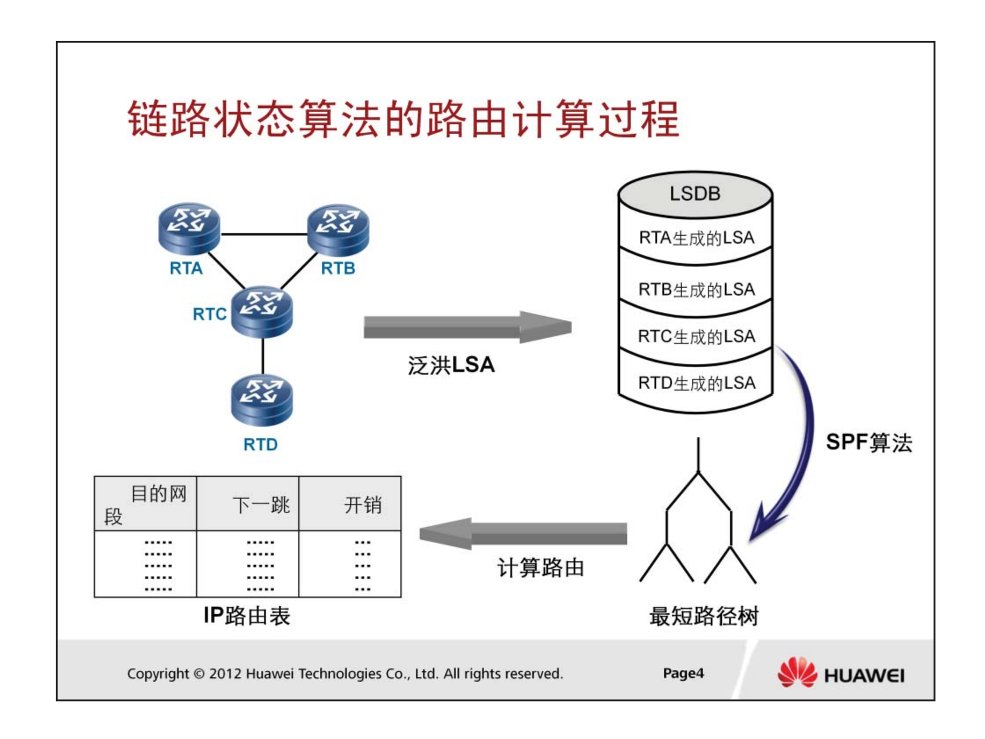
OSPF路由器使用组播和单播收发协议数据，因此占用的网络资源很小。

支持多条等值路由:

当到达目的地的等开销路径有多条时，流量被均衡地分担在这些等开销路径上。

支持协议报文的认证:

OSPF路由器之间交换的所有报文都被验证。



OSPF最显著的特点是使用链路状态算法，区别于早先的路由协议使用的距离矢量算法，因此，本文首先介绍链路状态算法的路由计算基本过程。

每个路由器通过泛洪链路状态通告（LSA）向外发布本地链路状态信息（例如使能OSPF的端口，可到达的邻居以及相邻的网段等等）。

每一个路由器通过收集其它路由器发布的链路状态通告以及自身生成的本地链路状态通告，形成一个链路状态数据库（LSDB）。LSDB描述了路由域内详细的网络拓扑结构。

所有路由器上的链路状态数据库是相同的。

通过LSDB，每台路由器计算一个以自己为根，以网络中其它节点为叶的最短路径树。

通过每台路由器计算的最短路径树得出了到网络中其它节点的路由表。

基本概念

自治系统（Autonomous System）：

- 一个自治系统是指使用同一种路由协议交换路由信息的一组路由器。

Router ID：

- 用于在自治系统中唯一标识一台运行OSPF的路由器的32位整数，每个运行OSPF的路由器都有一个Router ID。

 Page5

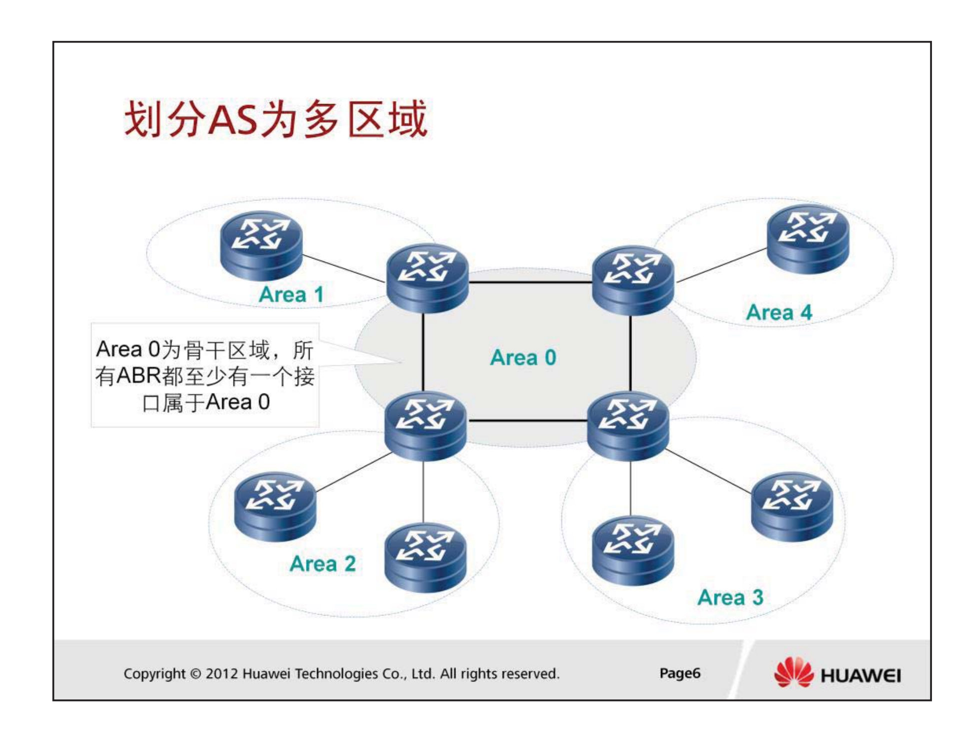


在OSPF中，有两个基本的概念需要介绍，一个是自治系统，或者说一个OSPF路由域；一个是Router ID。

在OSPF课程中，自治系统（Autonomous System）是指使用同一种路由协议交换路由信息的一组路由器，简称AS。

由于LSDB描述的是整个网络的拓扑结构，包括网络内所有的路由器，所以网络内每个路由器都需要有一个唯一的标识，用于在LSDB中标识自己。Router ID就是这样一个用于在自治系统中唯一标识一台运行OSPF的路由器的32位整数。每个运行OSPF的路由器都有一个Router ID。

Router ID的格式和IP地址的格式是一样的，推荐使用路由器Loopback0的IP地址做为路由器的Router ID。



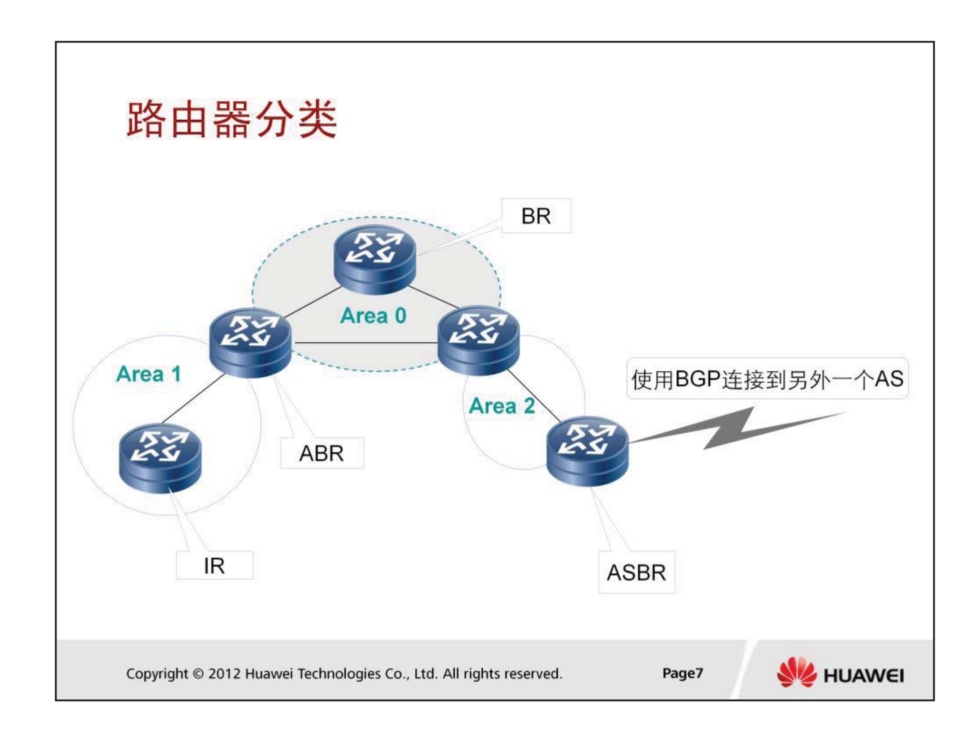
区域是一组网段的集合。

OSPF支持将一组网段组合在一起，这样的一个组合称为一个区域，即区域是一组网段的集合。

划分区域可以缩小LSDB规模，减少网络流量。

区域内的详细拓扑信息不向其他区域发送，区域间传递的是抽象的路由信息，而不是详细的描述拓扑结构的链路状态信息。每个区域都有自己的LSDB，不同区域的LSDB是不同的。路由器会为每一个自己所连接到的区域维护一个单独的LSDB。由于详细链路状态信息不会被发布到区域以外，因此LSDB的规模大大缩小了。

Area 0为骨干区域，骨干区域负责在非骨干区域之间发布由区域边界路由器汇总的路由信息（并非详细的链路状态信息），为了避免区域间路由环路，非骨干区域之间不允许直接相互发布区域间路由信息。因此，所有区域边界路由器都至少有一个接口属于Area 0，即每个区域都必须连接到骨干区域。



内部路由器（Internal Router）：

内部路由器是指所有所连接的网段都在一个区域的路由器。属于同一个区域的IR维护相同的LSDB。

区域边界路由器（Area Border Router）：

区域边界路由器是指连接到多个区域的路由器。ABR为每一个所连接的区域维护一个LSDB。

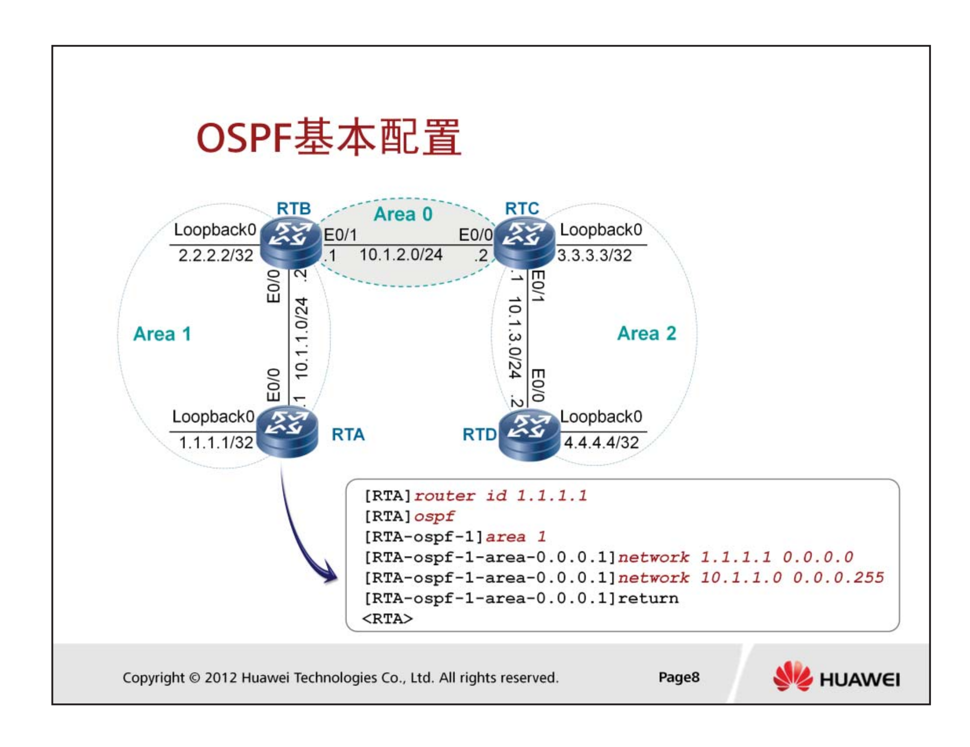
骨干路由器（Backbone Router）：

骨干路由器是指至少有一个端口（或者虚连接）连接到骨干区域的路由器。包括所有的ABR和所有端口都在骨干区域的路由器。

AS边界路由器（AS Boundary Router）：

AS边界路由器是指和其他AS中的路由器交换路由信息的路由器，这种路由器向整个AS通告AS外部路由信息。

AS边界路由器可以是内部路由器IR，或者是ABR，可以属于骨干区域也可以不属于骨干区域。



物理拓扑描述：

网络中共有四个路由器，每个路由器使用Loopback0接口的IP地址做为Router ID。整个路由域分为三个区域。RTB和RTC做为ABR。

此处省略端口和IP地址的配置。

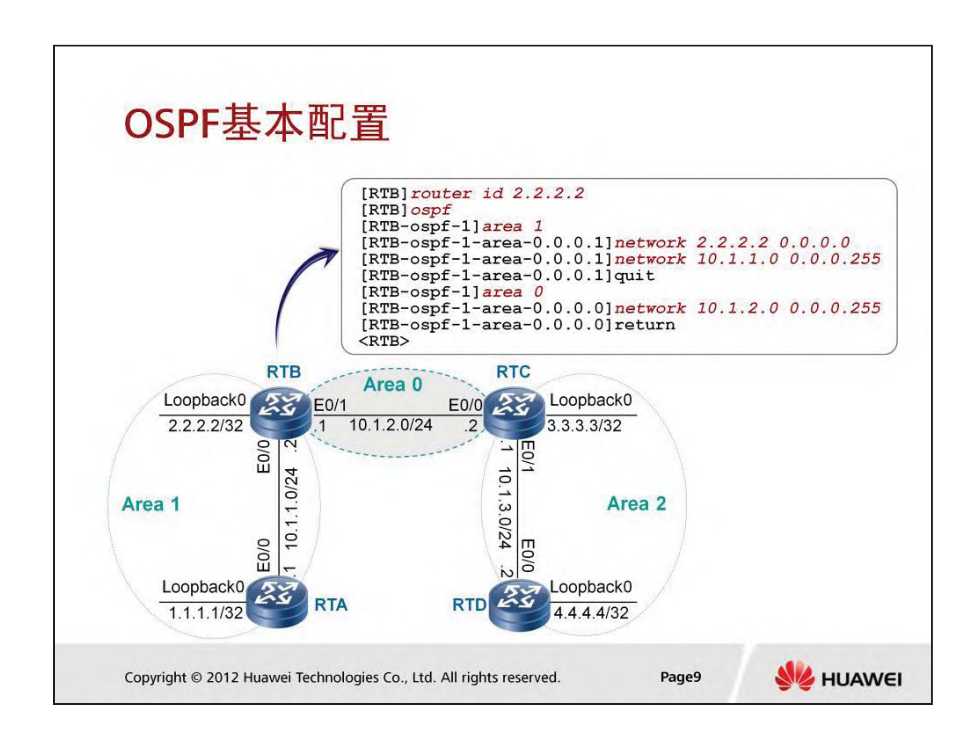
OSPF基本配置包括：

router id router-id ： 指定此路由器的Router ID。如果不手动指定Router ID，则OSPF自动使用Loopback接口中最大的IP地址做为Router ID，如果没有配置Loopback接口，则使用物理接口中最大的IP地址做为RouterID；

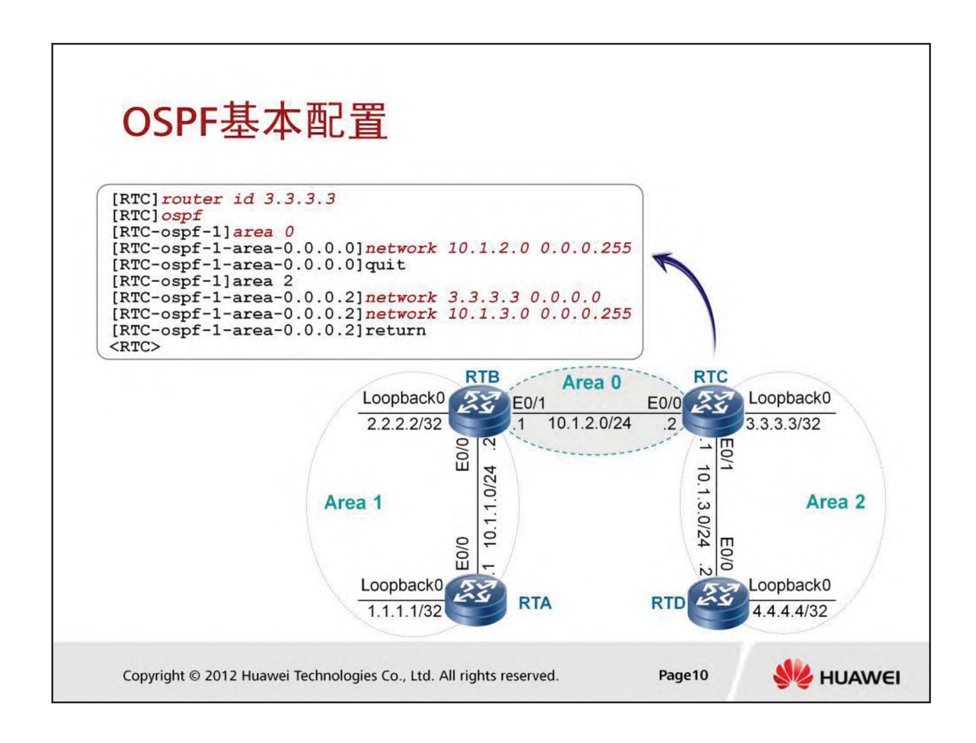
ospf process-id ： 开启OSPF。OSPF支持多进程，如果不指定进程号，默认使用进程号码1；

area area-id ： 进入区域视图；

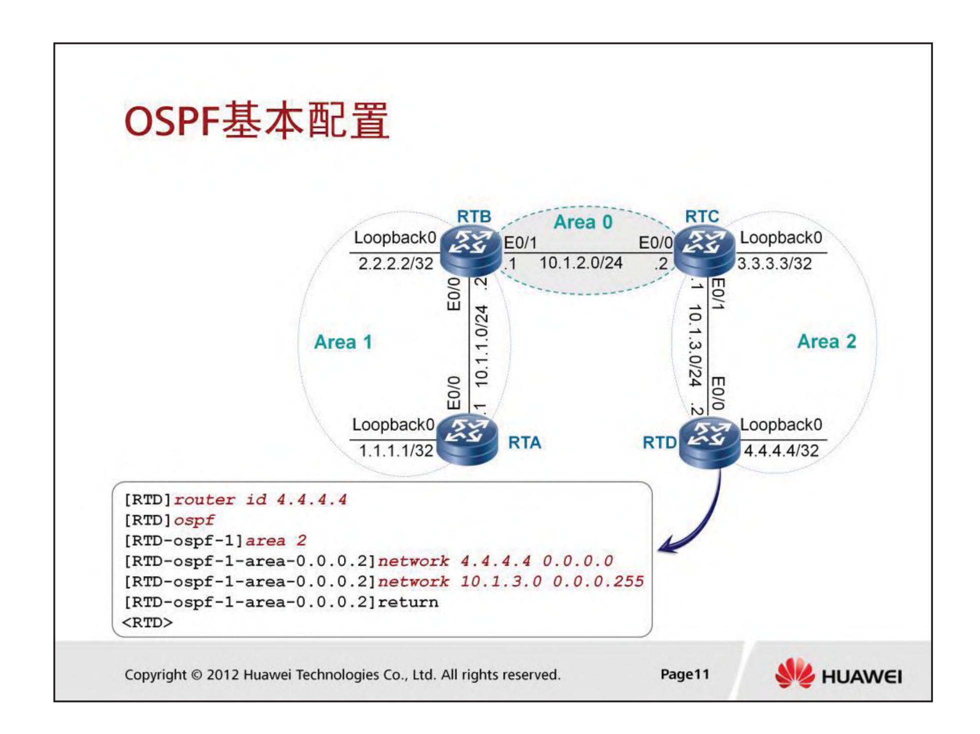
network ip-address wildcard ： 指定接口所在的网段地址，指定网段时，要使用该网段网络掩码的反码。



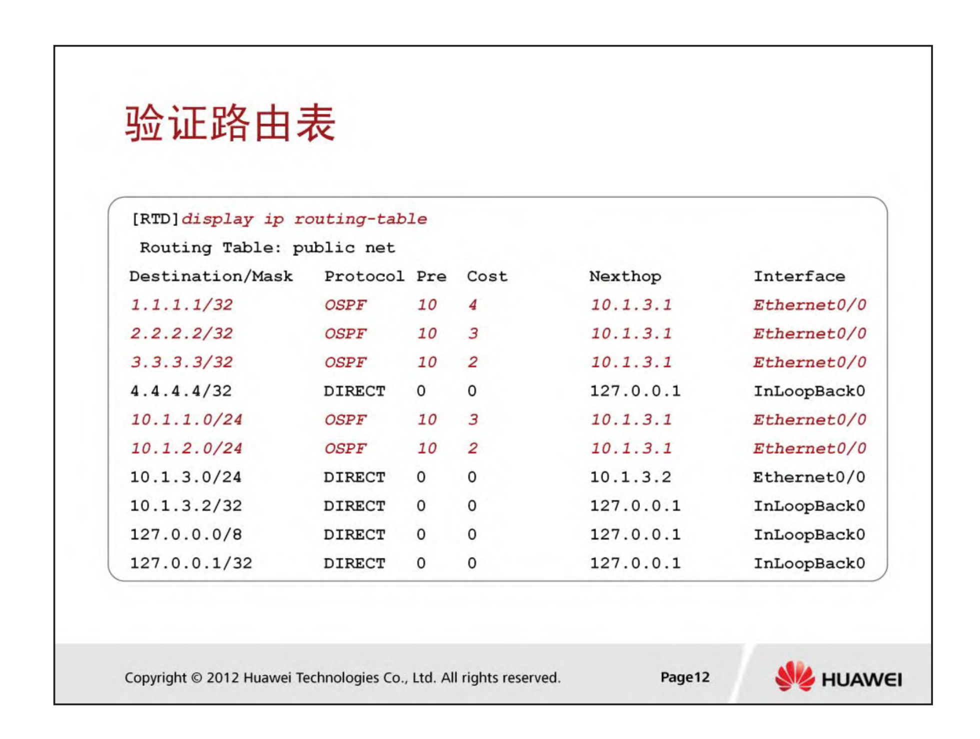
在RTB上需要配置两个区域，一个是骨干区域，一个为非骨干区域。Loopback0接口地址只在一个区域内宣告即可。



在RTC上需要配置两个区域，一个为骨干区域，一个为非骨干区域。



在RTD上只有一个区域（Area 2）。



路由表中有五条路由条目通过OSPF学到。



问 题

请描述链路状态算法的基本计算过程?

什么是OSPF区域?

OSPF基本配置包括哪些步骤?

Page13



请描述链路状态算法的基本计算过程。

每个路由器向外发布本地链路状态信息，并收集其它路由器发布的链路状态信息，形成一个描述网络拓扑结构的链路状态数据库，通过此数据库使用最短路径优先算法计算一个最短路径树，最短路径树给出了到网络中每个节点的路由。

什么是OSPF区域?

一个OSPF区域是一组网段集合。

OSPF基本配置包括哪些步骤?

开启OSPF进程，创建OSPF区域，指定每个区域中所包含的网段。


谢谢
www.huawei.com


理解OSPF邻居与邻接关系
www.huawei.com
Copyright © 2012Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍OSPF邻居和邻接的概念。

为了交换链路状态信息以及路由信息，OSPF路由器之间首先要建立邻接关系。

Page1





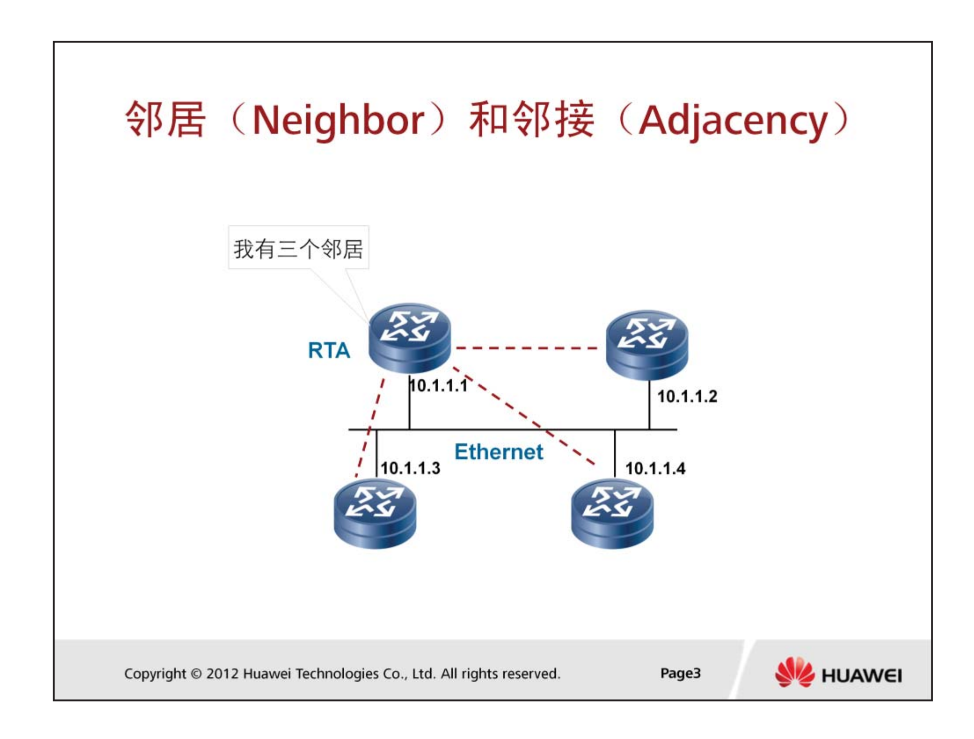
培训目标

学完本课程后，您应该能：

- 理解OSPF邻居和邻接的概念
- 理解OSPF中DR和BDR的概念
- 理解DR和BDR的选举

Page2





OSPF是一个动态路由协议，运行OSPF的路由器之间需要交换链路状态信息和路由信息，在交换这些信息之前首先需要建立邻接关系。

邻居路由器（Neighbor）：

有端口连接到同一个网段的两个路由器就是邻居路由器。

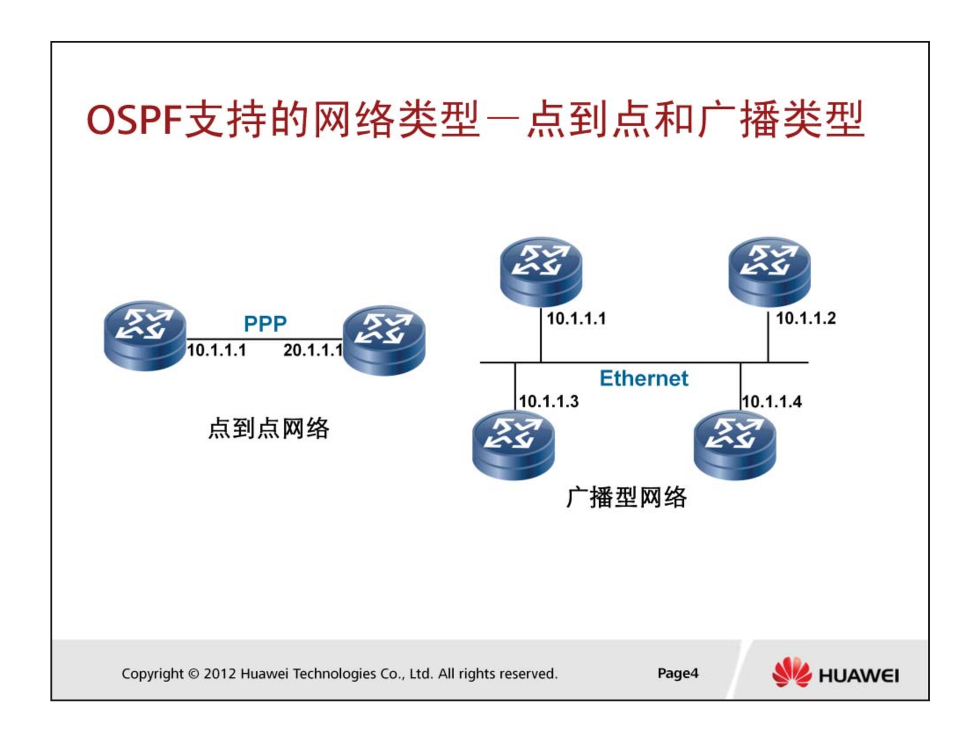
邻居关系由OSPF的Hello协议维护。

邻接（Adjacency）：

从邻居关系中选出的为了交换路由信息而形成的关系。

并非所有的邻居关系都可以成为邻接关系，不同的网络类型，是否建立邻接关系的规则也不同。

本例中，RTA有三个邻居。



前面提到，并非所有的邻居关系都可以形成邻接关系进而交换链路状态信息以及路由信息，是否建立邻接关系与网络类型有关。所谓网络类型是指运行OSPF网段的二层链路类型。

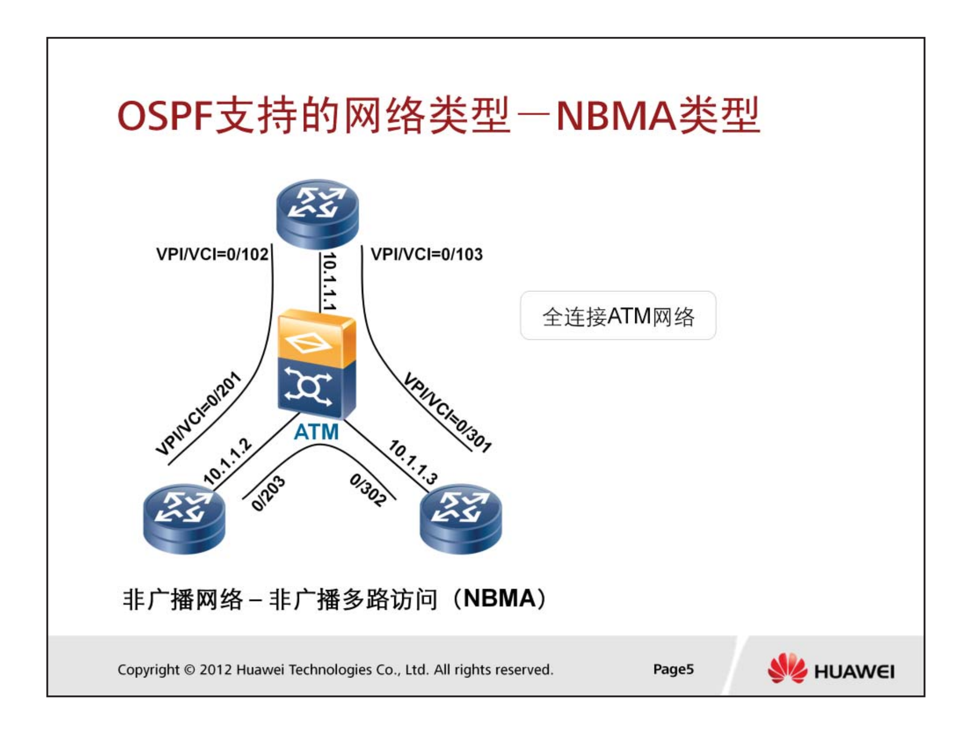
OSPF定义了四种网络类型，分别是点到点网络，广播型网络，NBMA网络和点到多点网络。

点到点网络是指只把两台路由器直接相连的网络。

一个运行PPP的64K串行线路就是一个点到点网络的例子。

广播型网络是指支持两台以上路由器，并且具有广播能力的网络。

一个含有四台路由器的以太网就是一个广播型网络的例子。



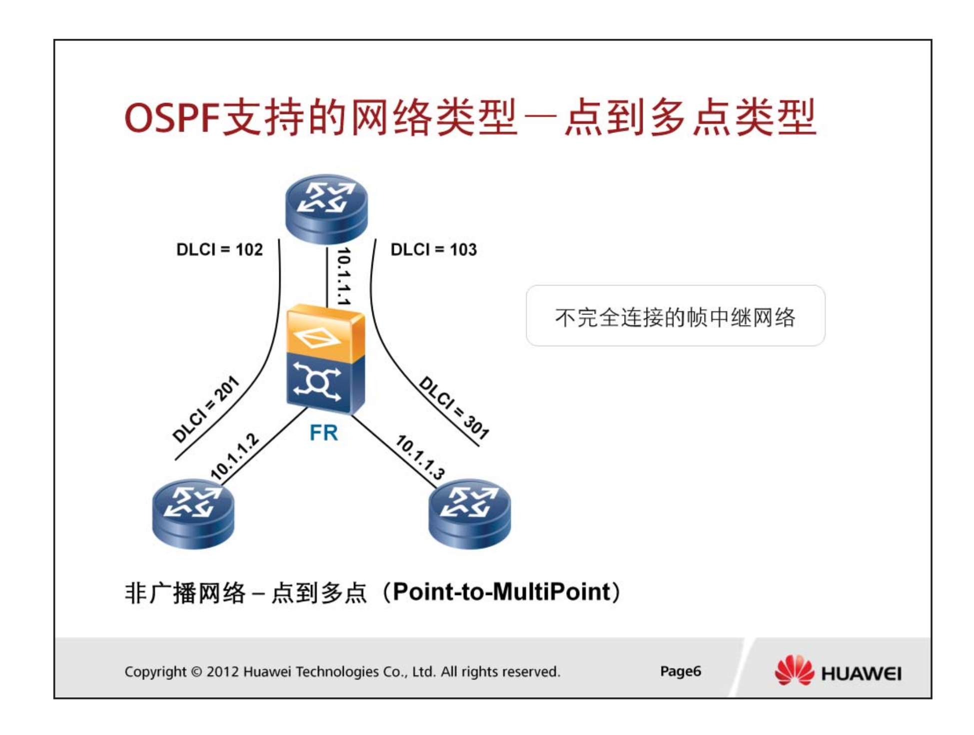
非广播网络是指支持两台以上路由器互连，但是不具有广播能力的网络。

在非广播网络上，OSPF有两种运行方式，非广播多路访问和点到多点。

非广播多路访问（NBMA）：

在NBMA网络上，OSPF模拟在广播型网络上的操作，但是每个路由器的邻居需要手动配置。

NBMA方式要求网络中的路由器组成全连接。例如，使用SVC进行通信的ATM网络。



点到多点：

将整个非广播网络看成是一组点到点网络。每个路由器的邻居可以使用底层协议例如反向地址解析协议（Inverse ARP）来发现。

对于不能组成全连接的网络应当使用点到多点方式，例如只使用PVC的不完全连接的帧中继网络。

常见链路层协议对应的默认网络类型

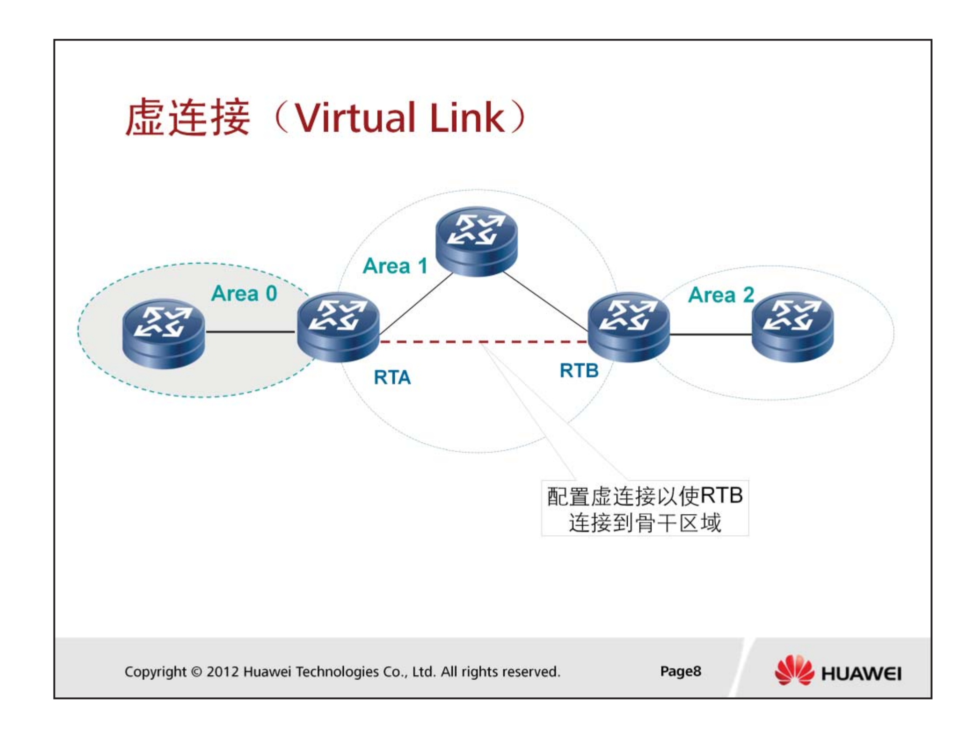
网络类型	常见链路层协议
Point-to-point	PPP链路；LAPB链路；HDLC链路
Broadcast	以太网链路
NBMA	帧中继链路；ATM链路

 Page7



此表列出了常见的链路层协议所对应的默认网络类型。

点到多点（Point-to-MultiPoint）网络类型不是一种默认的网络类型。



除了上述四种物理网络类型之外，还有一种虚拟链路类型 – 虚连接。

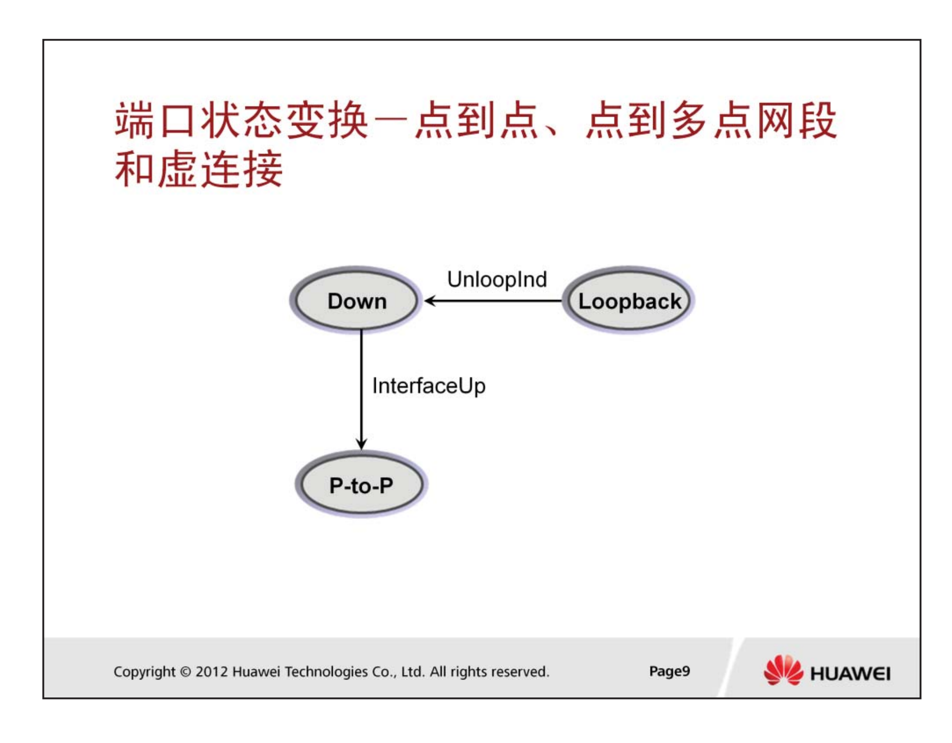
骨干区域必须是连续的，但在物理上不连续的时候，可以使用虚连接使骨干区域在逻辑上连续。

虚连接可以在任意两个区域边界路由器上建立，但是要求这两个区域边界路由器都有端口连接到一个共同的非骨干区域。这个非骨干区域成为Transit区域。

如上图所示，RTB做为一个ABR没有物理连接到骨干区域，此时可以在RTA和RTB之间配置一条虚拟链路，使RTB连接到骨干区域。Area 1是此虚拟连接的Transit区域。

虚连接技术虽然理论上使骨干区域可以在物理上不连续，但在实际组网时是不推荐的。

虚连接是属于骨干区域（Area 0）的一条虚拟链路。



各种状态的解释如下：

Down：

这是端口的初始状态，在该状态下，底层协议显示该端口不可用，所有定时器被关闭。

Loopback：

此状态表示端口被环回。在该状态下的端口被通告为一个Stub网段。

Point-to-point（P-to-P）：

在此状态下，端口是可用的，而且端口是连接到点到点、点到多点或者虚连接，此状态下的端口试图与邻居建立邻接关系，并以HelloInterval的间隔发送Hello报文。

各种事件解释如下：

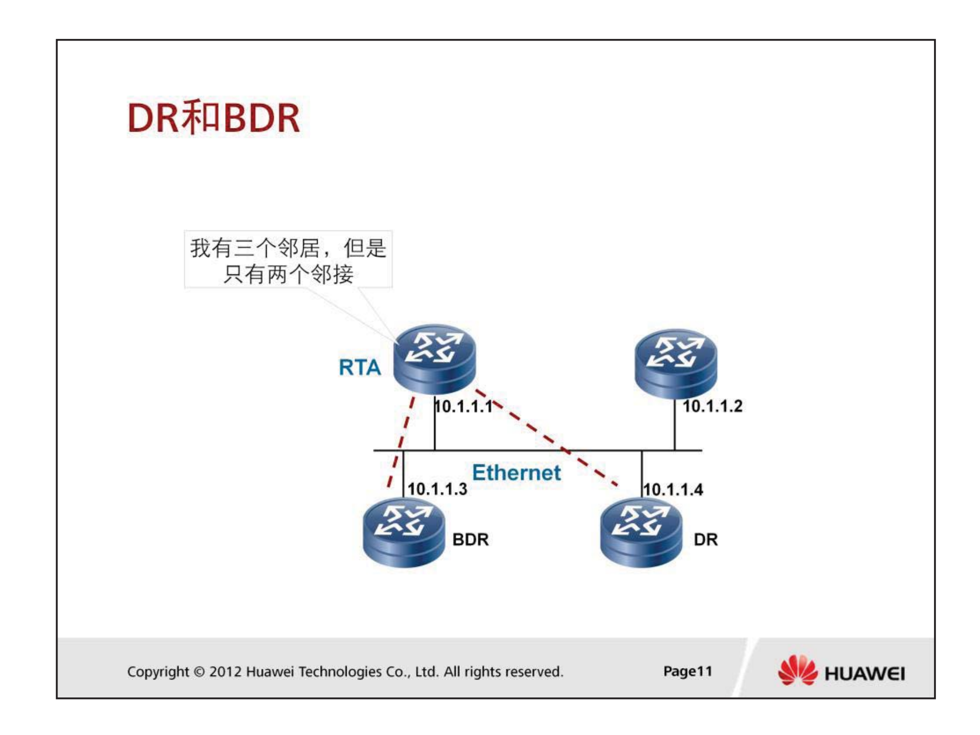
UnloopInd：

Unloopback Indication，表示端口解除环回状态。处于Loopback状态下的端口如果收到此事件，则进入Down状态。

InterfaceUp：

端口的链路层协议变成可用状态，即常说的链路层Up。由于不需要选举DR和BDR，因此点到点、点到多点网段以及虚连接的端口状态变换比较

简单，在Down状态下收到InterfaceUp事件后，转为Point-to-point（P-to-P）状态，此状态即为稳定工作状态。



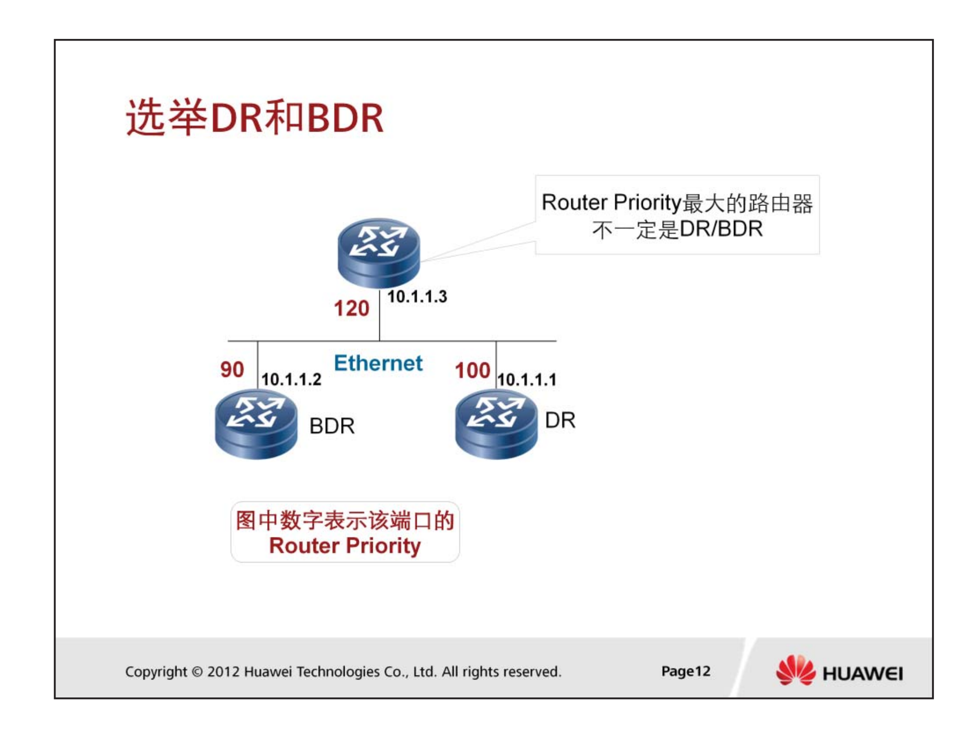
每一个含有至少两个路由器的广播型网络和NBMA网络都有一个指定路由器（Designated Router，DR）和备份指定路由器（Backup Designated Router，BDR）。

DR和BDR的作用:

1. 减少邻接关系的数量，从而减少链路状态信息以及路由信息的交换次数，这样可以节省带宽，减少路由器硬件的负担。一个既不是DR也不是BDR的路由器只与DR和BDR形成邻接关系并交换链路状态信息以及路由信息，这样就大大减少了大型广播型网络和NBMA网络中的邻接关系数量。

本例中，虽然RTA有三个邻居，但是只形成两个邻接关系。

2. 在描述拓扑的LSDB中，一个NBMA网段或者广播型网段是由单独一条LSA来描述的，这条LSA是由该网段上的DR产生的。



DR和BDR由OSPF的Hello协议选举，选举是根据端口的路由器优先级（Router Priority）进行的。

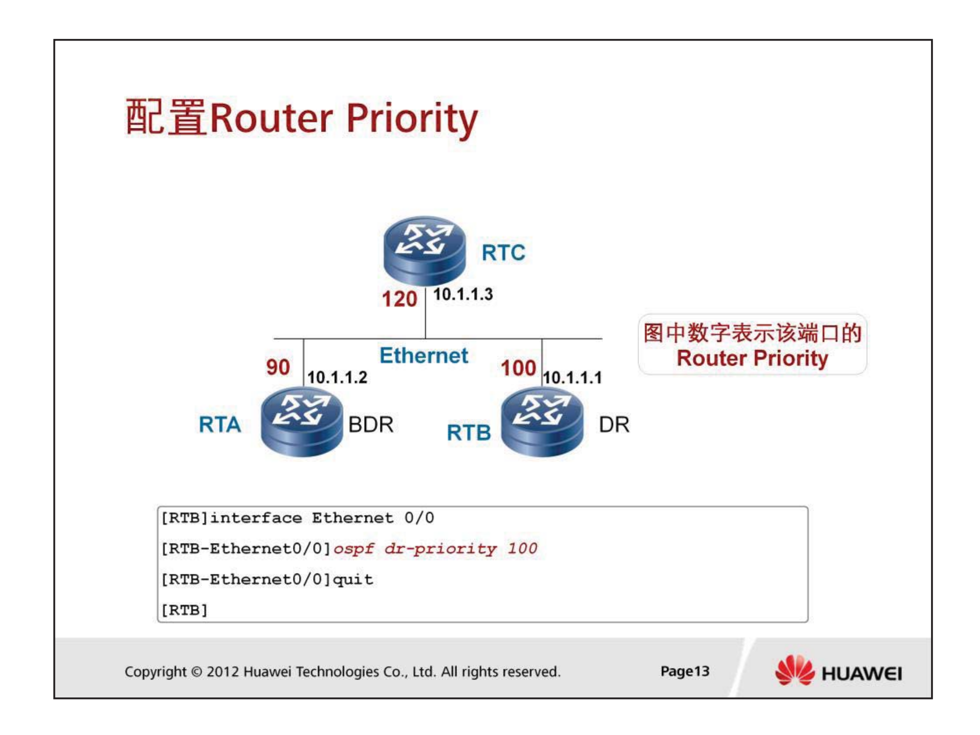
如果Router Priority被设置为0，那么该路由器将不允许被选举成DR或者BDR。

Router Priority越大越优先。如果相同，Router ID大者优先。

但是为了维护网络上邻接关系的稳定性，如果网络中已经存在DR和BDR，则新添加进该网段的路由器不会成为DR和BDR，不管该路由器的Router Priority是否最大。

如果当前DR故障，当前BDR自动成为新的DR，网络中重新选举BDR；如果当前BDR故障，则DR不变，重新选举BDR。

这种选举机制的目的是为了保持邻接关系的稳定，减小拓扑结构的改变对邻接关系的影响。

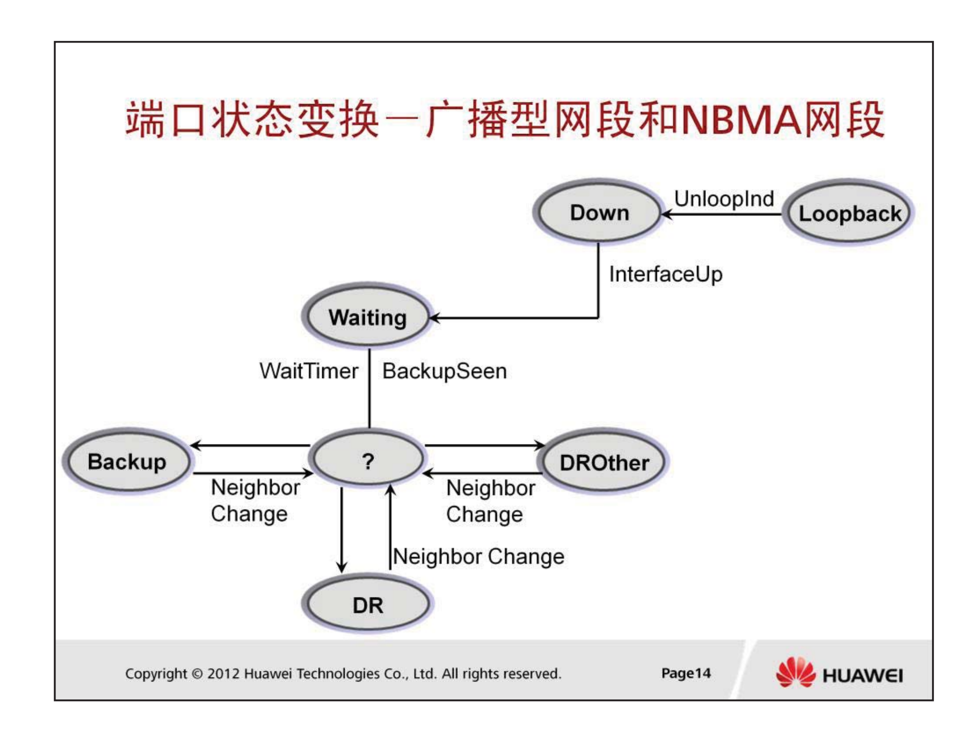


ospf dr-priority value：修改端口的Router Priority。

Router Priority的取值范围是0～255，默认值为1。

如果两台路由器Router Priority值相同，则比较Router ID，Router ID大的更优先。

如果修改了Router Priority，需要重启ospf进程才能重新参与选举DR和BDR。



相关状态和事件解释如下：

Waiting：

在此状态下，路由器通过监听接收到的Hello报文检测网络中是否已经有DR和BDR。在此状态下的路由器不可以参与选举DR和BDR。

Backup：

在此状态下，该路由器成为所连接网络上的BDR，并与网段中所有的其他路由器建立邻接关系。

DR：

在此状态下，该路由器成为所连接网络上的DR，并与网段中所有的其他路由器建立邻接关系。

DROther：

该路由器连接到一个广播型网段或者NBMA网段，而且该路由器不是一个DR或者BDR。此状态下的路由器与DR和BDR形成邻接关系并交换路由信息。

BackupSeen：

路由器已经检测到网络上是否存在BDR。

一个OSPF路由器在广播型网段和NBMA网段上选举DR和BDR之前，首先

会等待一段时间（RouterDeadInterval），在这段时间里检测网络上是否已经存在DR和BDR，如果已经有DR和BDR，则不启动选举过程，直接进入DROther状态。因此，网络上Router Priority最大的路由器不一定是DR，Router Priority第二大的路由器也不一定是BDR。

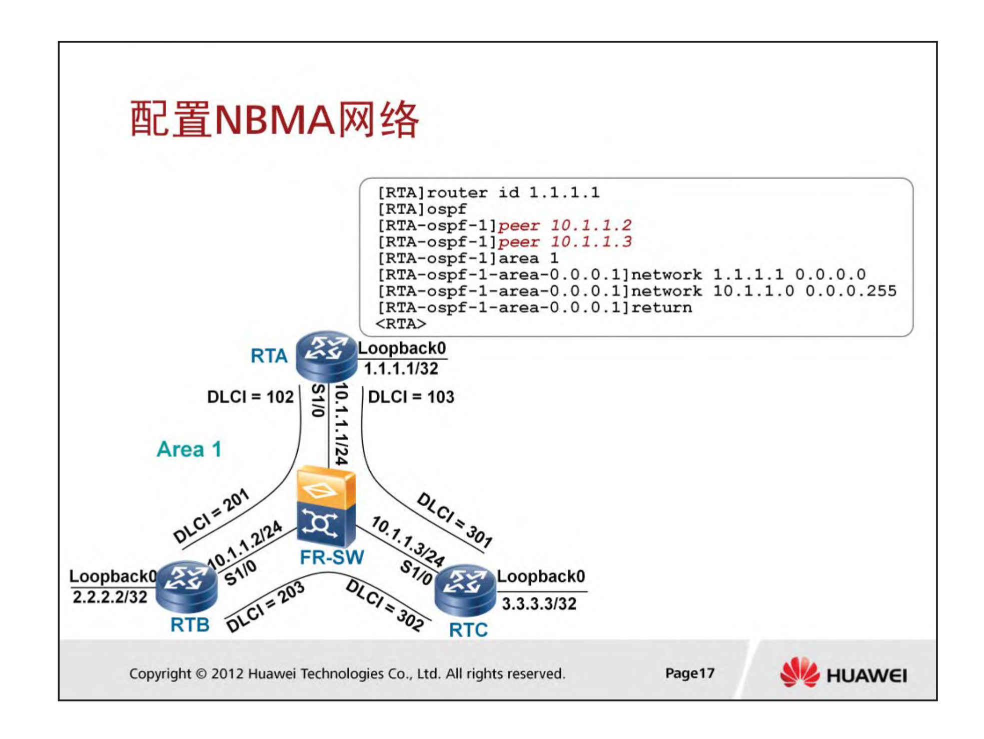
是否和邻居建立邻接关系

网络类型	是否和邻居建立邻接关系
Point-to-point	总是和邻居建立邻接关系
Point-to-MultiPoint	总是和邻居建立邻接关系
Virtual link	总是和邻居建立邻接关系
Broadcast NBMA	DR总是和其他所有路由器包括BDR建立邻接关系； BDR总是和其他所有路由器包括DR建立邻接关系； 处于DROther状态的路由器只与DR和BDR建立邻接关系

 Page16 

点到点，点到多点以及虚连接链路上，总是和邻居建立邻接关系。

在NBMA和广播型网段上，邻接关系的数量会比邻居关系的数量少一些。



物理拓扑描述:

三台路由器通过帧中继交换机组成全连接。把所有网段配置在Area 1中
。

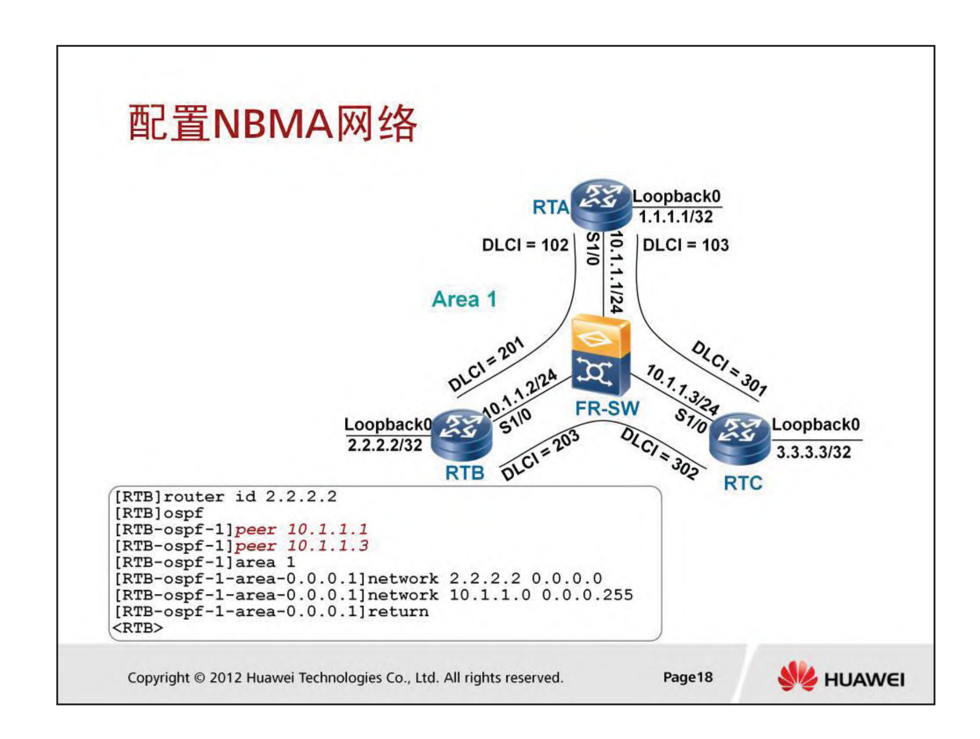
NBMA网络中不需要配置反向ARP，邻居需要手动指定。

在RTA上配置两个邻居，10.1.1.2和10.1.1.3。

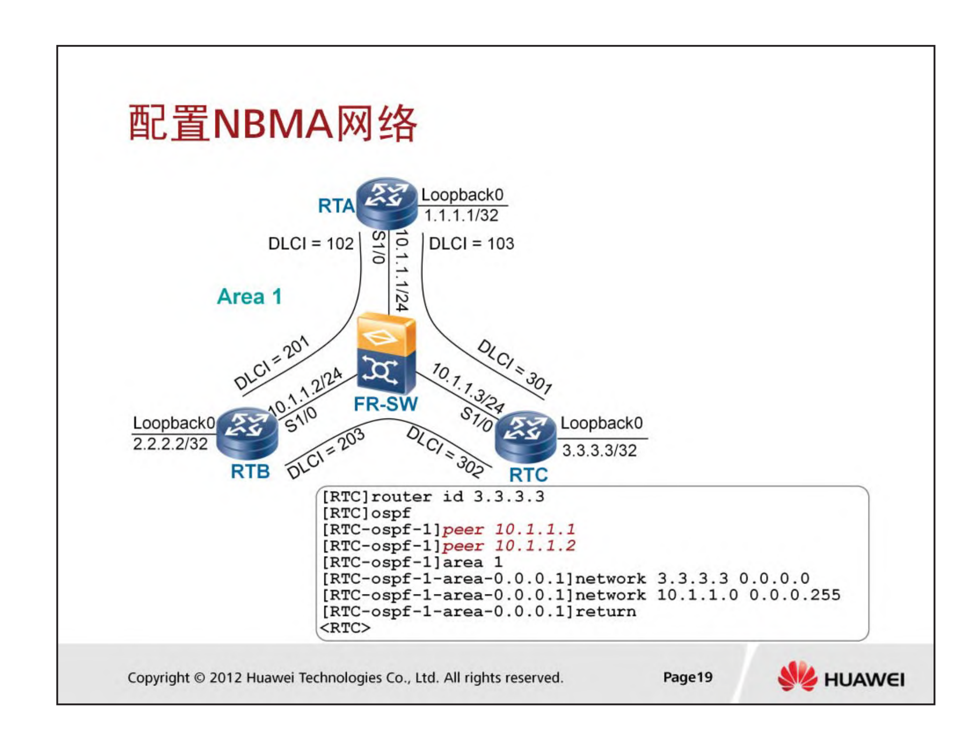
指定邻居的时候使用该邻居在该网段上的IP地址来标识。

peer ip-address [dr-priority dr-priority-number]

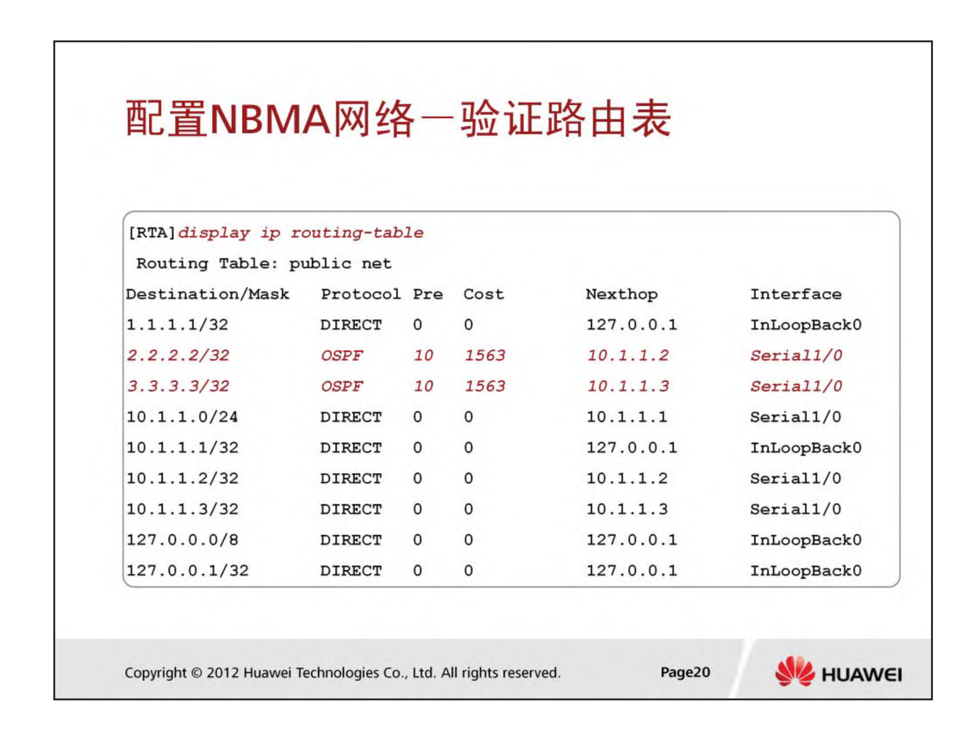
dr-priority-number：邻居的Router Priority，默认为1。



在RTB上配置两个邻居，10.1.1.1和10.1.1.3。



在RTC上配置两个邻居，10.1.1.1和10.1.1.2。



通过OSPF学习到其它路由器Loopback0接口的路由。

配置NBMA网络－验证OSPF端口信息

```
[RTA]display ospf interface Serial 1/0

              OSPF Process 1 with Router ID 1.1.1.1
                      Interfaces

  Interface: 10.1.1.1 (Serial1/0)
    Cost: 1562 State: DROther    Type: NBMA
    Priority: 1
    Designated Router: 10.1.1.3
    Backup Designated Router: 10.1.1.2
    Timers: Hello 30, Dead 120, Poll 120, Retransmit 5, Transmit Delay 1
```

 Page21 HUAWEI

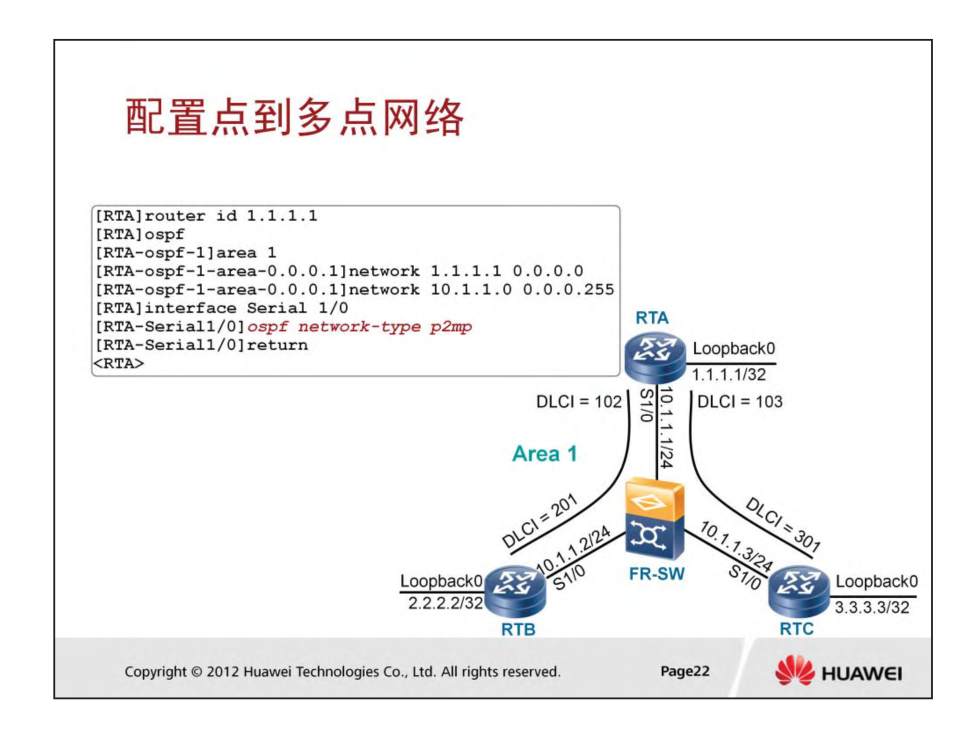
State: 端口状态。

Type: 该端口的OSPF网络类型。

Priority：该端口的Router Priority值，用于DR和BDR选举。

Designated Router: DR的端口IP地址。

Backup Designated Router: BDR的端口IP地址。



本例中：

RTA可以连接到其它两台路由器，但是RTB和RTC之间没有连接。

由于网络中的路由器不是全连接的，所以帧中继端口的OSPF网络类型需要手动指定为点到多点。

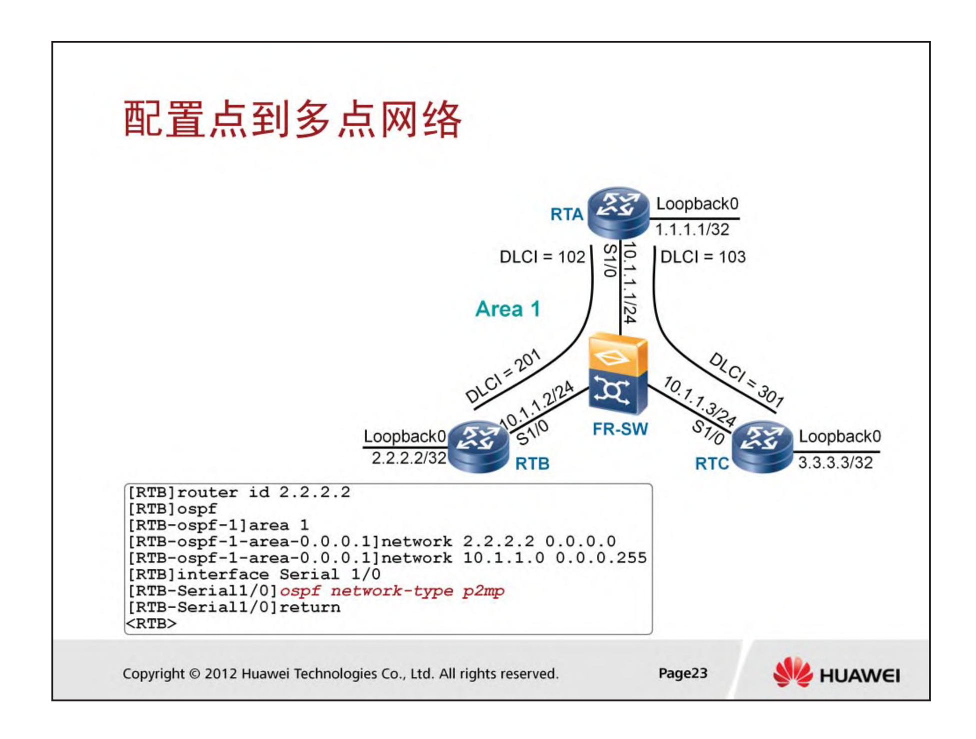
帧中继的反向ARP需要在端口上开启。

在RTA上，把所有网段配置在Area 1中。

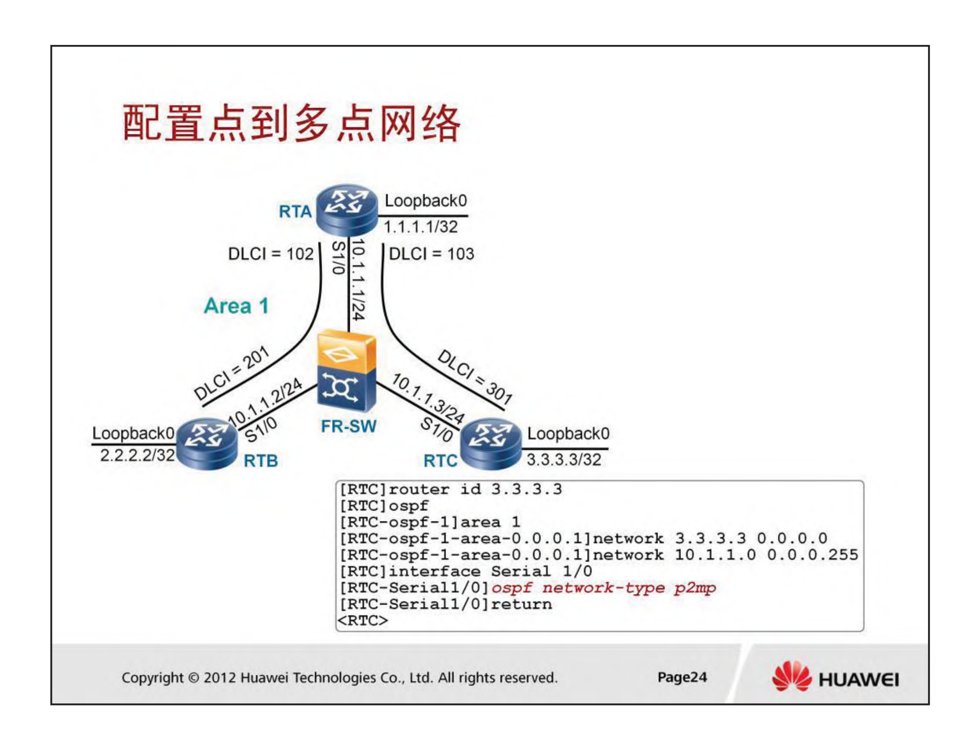
把Serial 1/0端口的网络类型手动修改成点到多点。

ospf network-type broadcast | nbma | p2mp | p2p

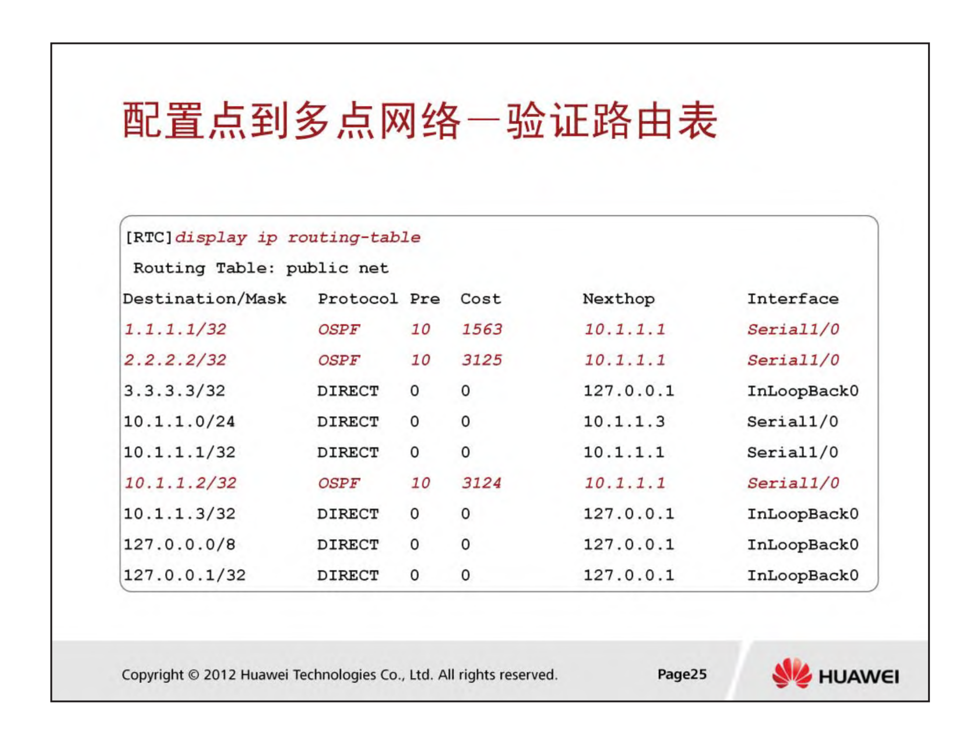
OSPF共有四种网络类型。



在RTB上，把所有网段配置在Area 1中，配置Serial 1/0的网络类型为点到多点。



在RTC上，把所有网段配置在Area 1中，配置Serial 1/0的网络类型为点到多点。



在路由表中，到两个Loopback0接口的路由通过OSPF学习，到RTB物理接口的路由也是通过OSPF学习。

配置点到多点网络—验证OSPF端口信息

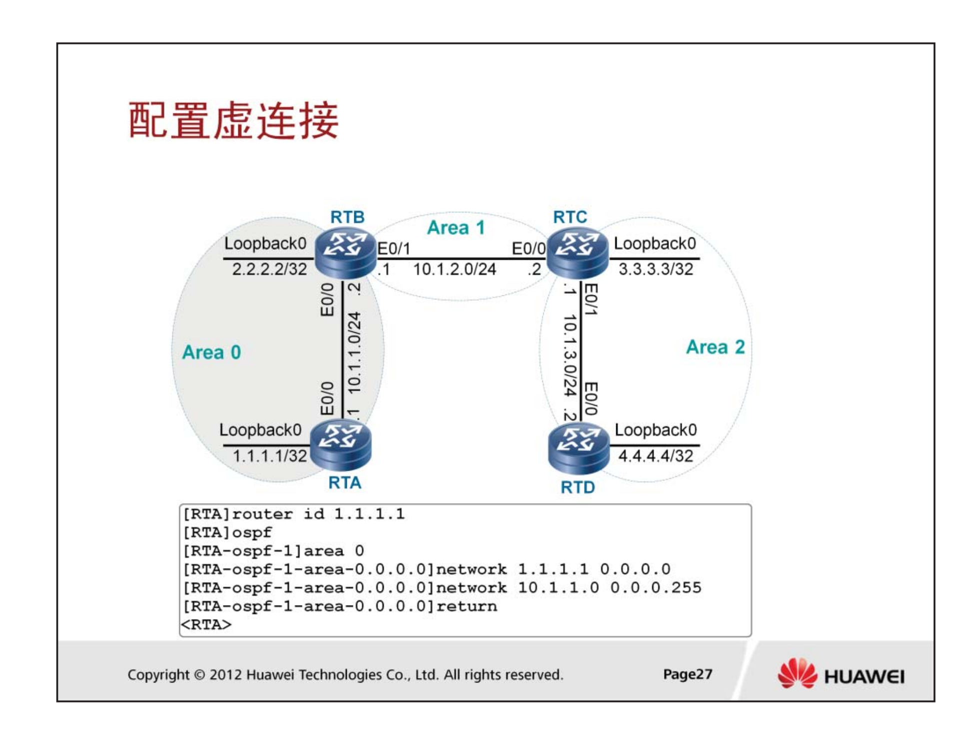
```
[RTC]display ospf interface Serial 1/0

                OSPF Process 1 with Router ID 3.3.3.3
                               Interfaces

    Interface: 10.1.1.3 (Serial1/0)
      Cost: 1562 State: PtoP    Type: PointToMultiPoint
      Priority: 1
      Timers: Hello 30, Dead 120, Poll 120, Retransmit 5, Transmit Delay 1
```

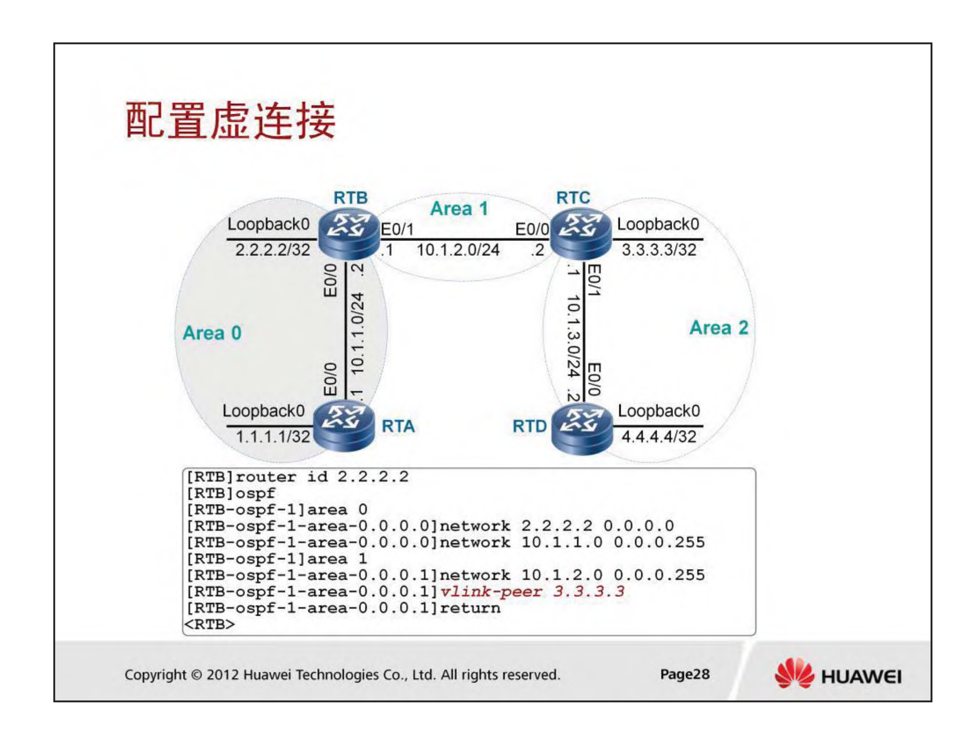

Page26
HUAWEI

OSPF网络类型为点到多点，点到多点网络类型的端口的端口稳定状态是Point-to-Point。



本例中:

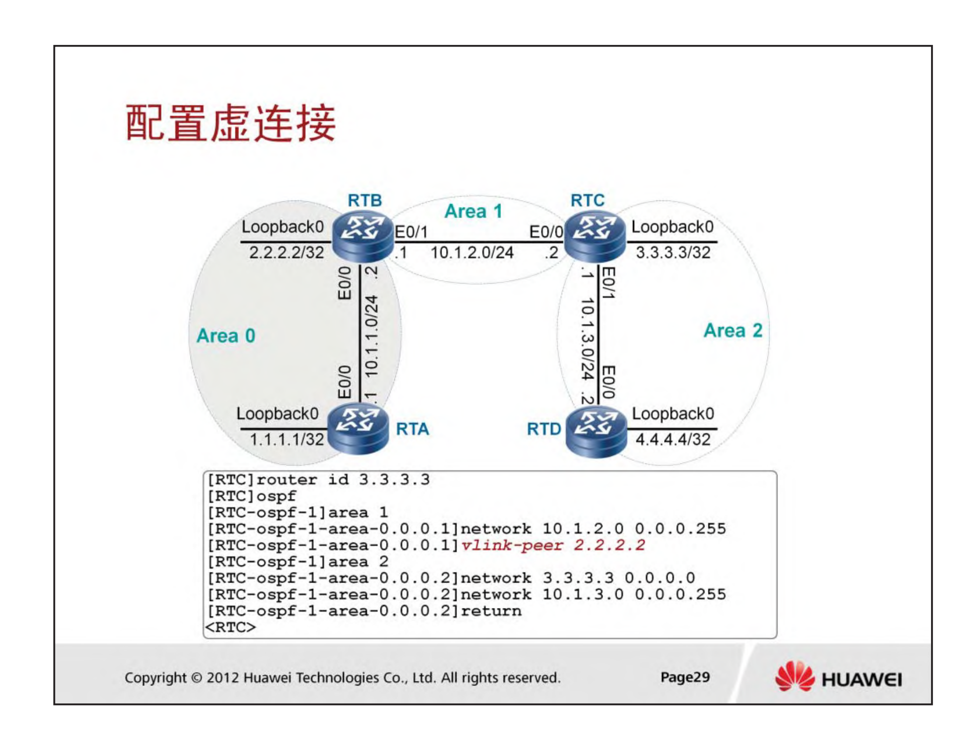
RTC是ABR，但是RTC没有连接到骨干区域。因此在RTB和RTC之间配置虚连接。Area 1是该虚连接的Transit区域。在RTA上，把所有网段配置在Area 0中。



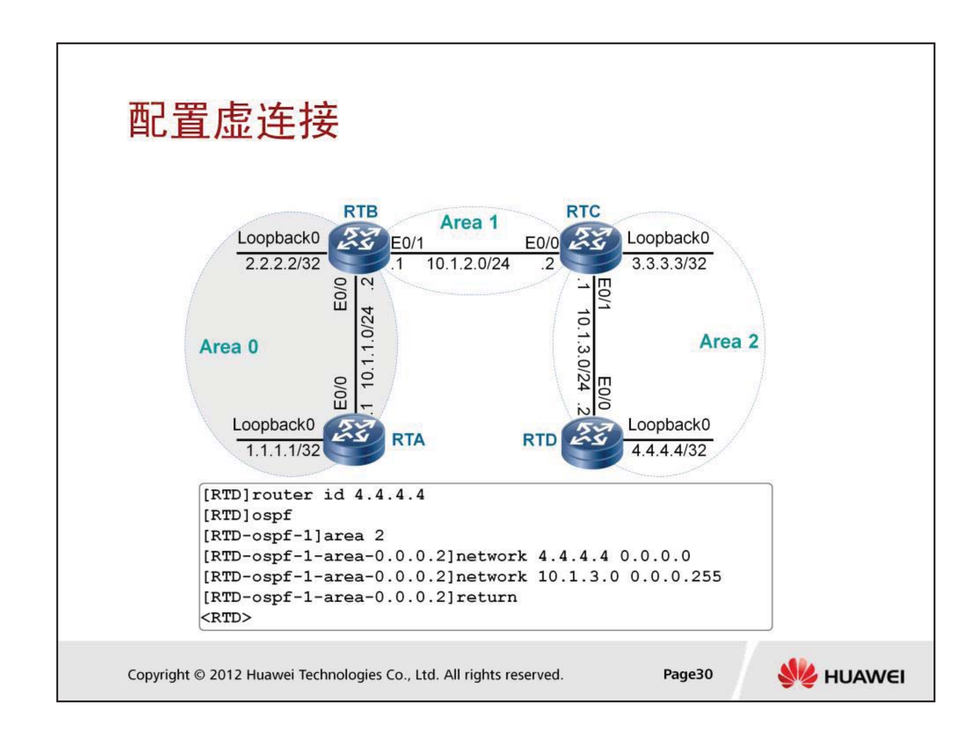
配置虚连接时，需要指定虚连接对端的Router ID。

虚连接在transit区域的区域视图下配置。

vlink-peer router-id： 配置虚拟连接时，使用对端Router ID表示对端路由器。



虚连接需要在两个ABR上配置。



把RTD上所有的网段配置在Area 2中。

配置虚连接—验证路由表

```
[RTD]display ip routing-table
 Routing Table: public net
Destination/Mask    Protocol Pre  Cost        Nexthop         Interface
1.1.1.1/32          OSPF     10   4           10.1.3.1        Ethernet0/0
2.2.2.2/32          OSPF     10   3           10.1.3.1        Ethernet0/0
3.3.3.3/32          OSPF     10   2           10.1.3.1        Ethernet0/0
4.4.4.4/32          DIRECT   0    0           127.0.0.1       InLoopBack0
10.1.1.0/24         OSPF     10   3           10.1.3.1        Ethernet0/0
10.1.2.0/24         OSPF     10   2           10.1.3.1        Ethernet0/0
10.1.3.0/24         DIRECT   0    0           10.1.3.2        Ethernet0/0
10.1.3.2/32         DIRECT   0    0           127.0.0.1       InLoopBack0
127.0.0.0/8         DIRECT   0    0           127.0.0.1       InLoopBack0
127.0.0.1/32        DIRECT   0    0           127.0.0.1       InLoopBack0
```

 Page31 HUAWEI

通过OSPF学习到五条路由。

配置虚连接—验证虚连接

```
[RTC]display ospf vlink

                OSPF Process 1 with Router ID 3.3.3.3
                          Virtual Links

 Virtual-link Neighbor-id  -> 2.2.2.2, State: Full
 Interface: 10.1.2.2 (Ethernet0/0)
   Cost: 1 State: PtoP    Type: Virtual
   Transit Area: 0.0.0.1
   Timers: Hello 10, Dead 40, Poll 0, Retransmit 5, Transmit Delay 1
```


HUAWEI

Neighbor-id:

虚连接邻居的Router ID。



问 题

邻居关系和邻接关系有什么区别?

OSPF支持的网络类型有哪些?

什么是DR和BDR?

Router Priority最大的一定是DR吗?

配置虚连接的时候如何表示对端路由器?

 Page33



Q：邻居关系和邻接关系有什么区别?

A：只要有端口连接到同一个网段的两个路由器就可以形成邻居关系，邻接关系是指可以交换链路状态信息以及路由信息的邻居关系，只有部分邻居关系可以形成邻接关系。

Q：OSPF支持的网络类型有哪些?

A：点到点网络，广播型网络，非广播多路访问网络，点到多点网络。

Q：什么是DR和BDR?

A：DR是广播型网段或者NBMA网段上的指定路由器，用于和其它路由器形成邻接关系，交换路由信息。

BDR是广播型网段或者NBMA网段上的备份指定路由器，用于和DR以及其他路由器形成邻接关系，交换路由信息。作为DR的备份路由器，当DR失效时，BDR将自动成为DR。

Q：Router Priority最大的一定是DR吗?

A：不一定，为了保持邻接关系的稳定性，拓扑结构的改变（不涉及当前DR和BDR）不会引起DR和BDR的重新选举。

Q：配置虚连接的时候如何表示对端路由器?

A：使用对端路由器的Router ID表示。


谢谢
www.huawei.com


OSPF协议报文和
链路状态通告
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍OSPF协议报文和链路状态通告。

课程内容包括协议报文头，报文类型，LSA类型等。

Page1





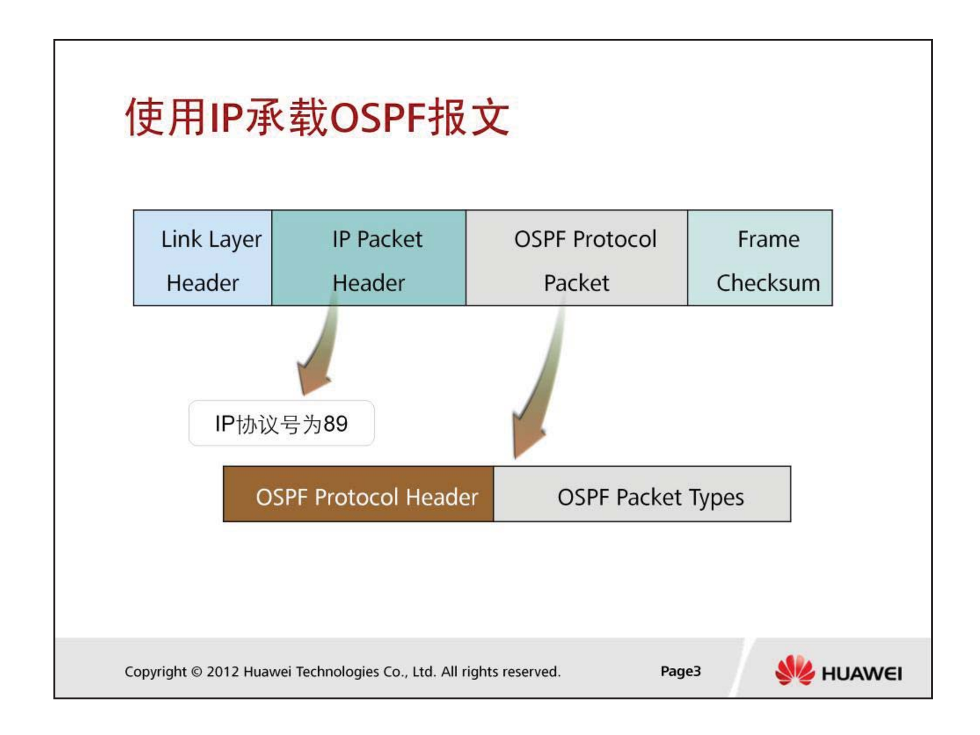
培训目标

学完本课程后，您应该能：

- 理解OSPF报文头和报文类型
- 理解链路状态通告类型

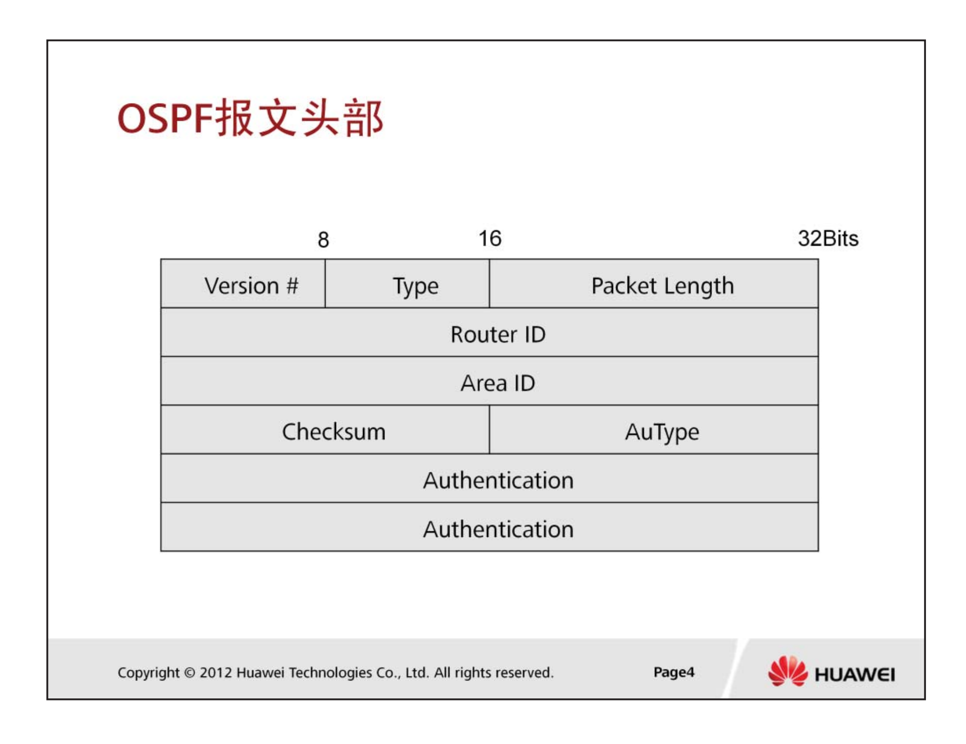
Page2





OSPF直接运行于IP协议之上，使用IP协议号89。

OSPF有五种报文类型，但是OSPF报文头部格式都是相同的。



所有的OSPF报文使用相同的OSPF报文头部。

Version #：

OSPF协议号，应当被设置成2。

Type：

OSPF报文类型，OSPF共有五种报文。

Packet length：

OSPF报文总长度，包括报文头部。单位是字节。

Router ID：

生成此报文的路由器的Router ID。

Area ID：

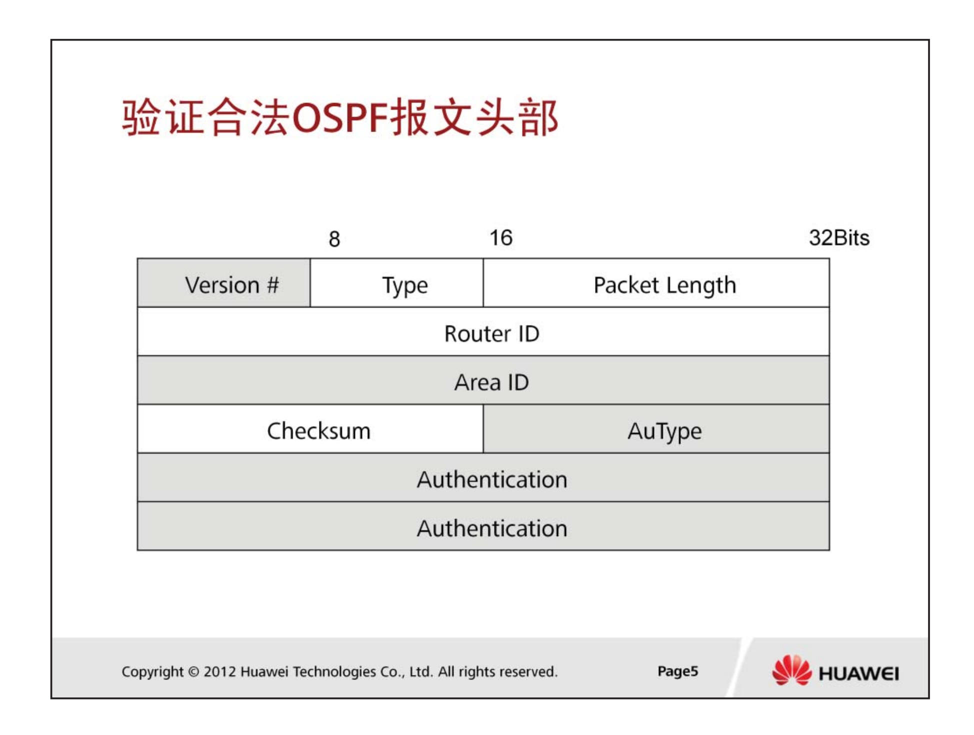
此报文需要被通告到的区域。

AuType：

验证此报文所应当使用的验证方法。

Authentication：

验证此报文时所需要的密码等信息。



验证一个OSPF报文头部是否合法包括：

1. Version必须为2；

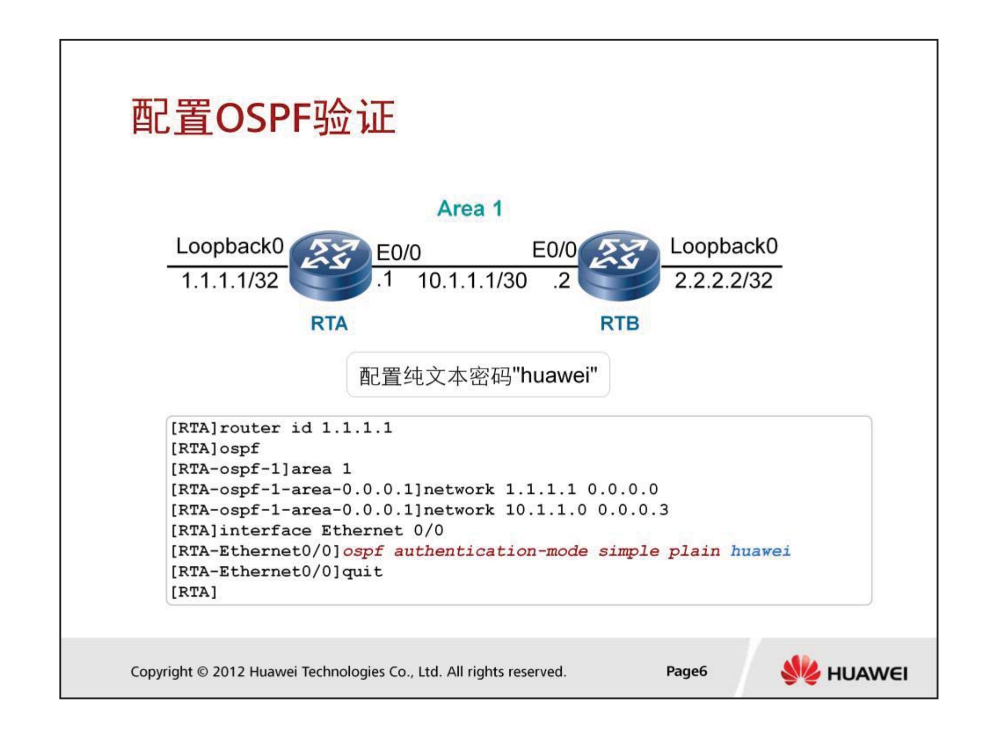
2. Area ID应当满足以下两种情况之一：a）和接收端口所属区域的Area ID一致；b）和接收端口所属区域的Area ID不一致，但是值为0，表示该报文属于骨干区域，而且是在一个虚连接上发送的；

3. AuType字段必须与该区域配置的Autype一致；

4. Authentication为验证信息，内容与AuType字段相关。

只有通过验证的OSPF报文才能被接受，否则将不能正常建立邻居关系。

VRP支持两种验证方式：区域验证方式和接口验证方式。当两种验证方式都存在时，优先使用接口验证方式。



OSPF报文的验证：

VRP中，OSPF支持区域验证和接口验证两种方式。

使用区域验证时，一个区域中所有路由器在该区域下的验证模式和口令必须一致；使用接口验证方式时，在相邻的路由器之间设置的验证模式和口令必须一致，优先级高于区域验证方式。

本例中，只有一个区域和两个路由器。在RTA上，配置验证方式为接口验证，配置验证模式为明文验证（simple），密码显示方式为明文(plain)，RTA和RTB之间的链路的密码为“huawei”。

如果验证方式为区域验证，则在区域视图下使用如下命令：

验证模式为明文验证：

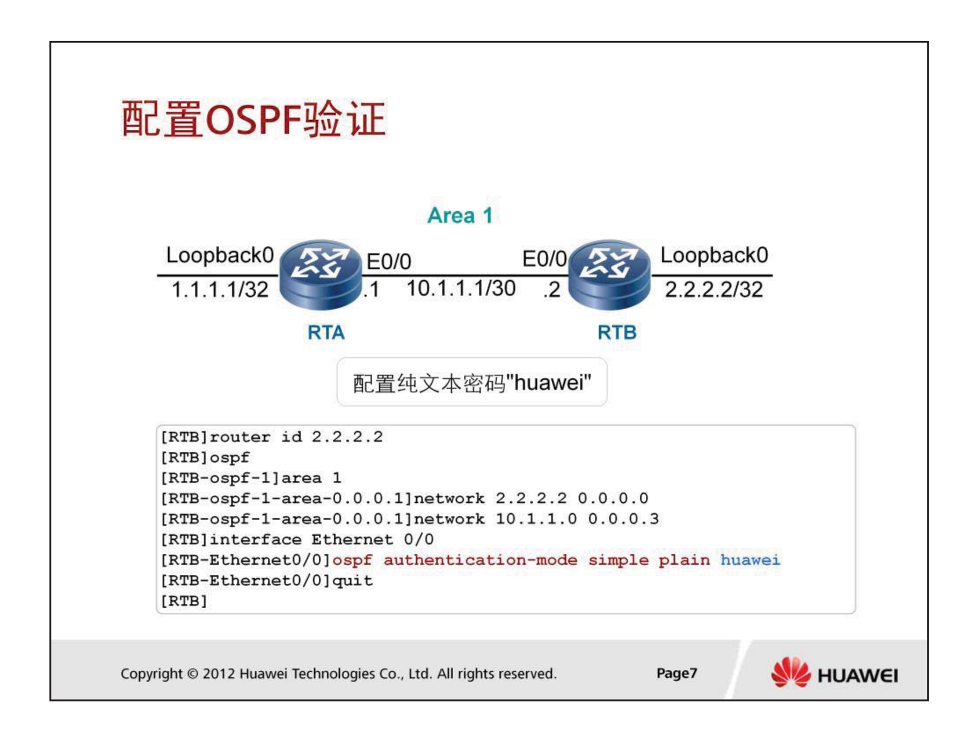
authentication-mode simple { [plain] plain-text | cipher cipher-text }

plain：密码显示方式为明文。

cipher：密码显示方式为密文。

验证模式为MD5验证：

authentication-mode md5 key-id { [plain] plain-text | cipher cipher-text }



在RTB上，在接口视图下指定验证模式为明文验证（simple），密码显示方式为明文(plain)，RTA和RTB之间的链路的密码为 “huawei”。

OSPF报文类型

Type	报文名称	报文功能
1	Hello	发现和维护邻居关系
2	Database Description	发送链路状态数据库摘要
3	Link State Request	请求特定的链路状态信息
4	Link State Update	发送详细的链路状态信息
5	Link State Ack	发送确认报文

 Page8 

OSPF共有五种报文类型。

Hello报文用于发现和维护邻居关系，在广播型网络和NBMA网络上Hello报文也用来选举DR和BDR。

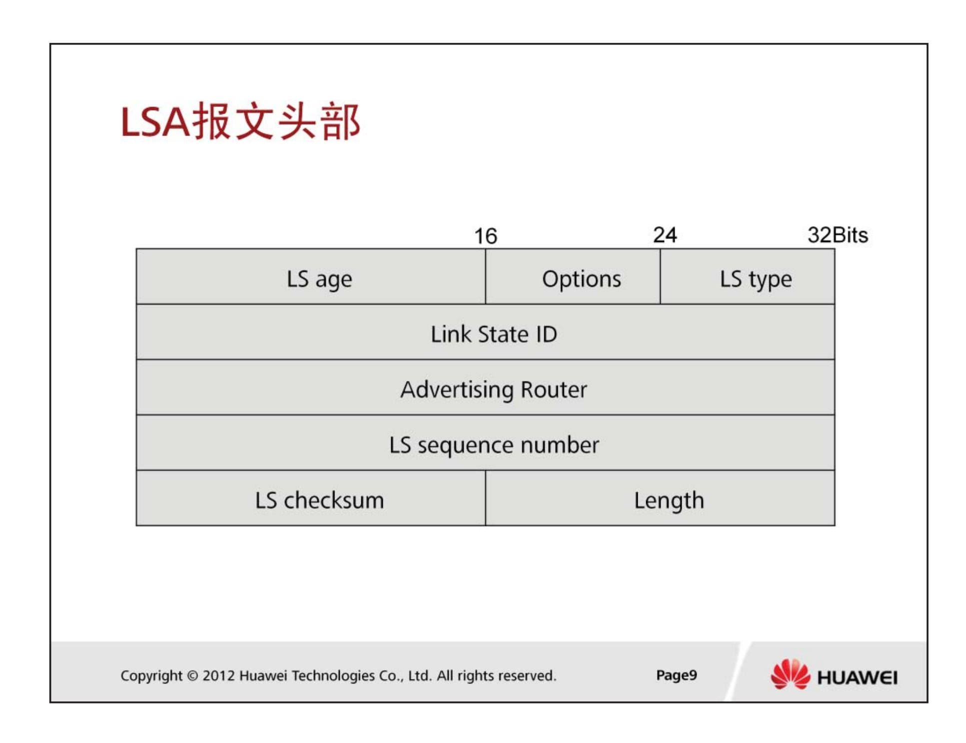
DD报文通过携带LSA头部信息来描述链路状态摘要信息。

LS Request报文用于发送下载LSA的请求信息，这些被请求的LSA是通过接收DD报文发现的，但是本路由器上没有的。

LS Update报文通过发送详细的LSA来同步链路状态数据库。

LS Ack报文通过泛洪确认信息确保路由信息的交换过程是可靠的。

除了Hello报文以外，其他所有报文只在建立了邻接关系的路由器之间发送。



除Hello报文外，其它的OSPF报文都携带LSA信息。

LS age：

此字段表示LSA已经生存的时间，单位是秒。

LS type：

此字段标识了LSA的格式和功能。常用的LSA类型有五种。

Link State ID：

此字段是该LSA所描述的那部分链路的标识。例如Router ID等。

Advertising Router：

此字段是产生此LSA的路由器的Router ID。

LS sequence number：

此字段用于检测旧的和重复的LSA。

LS type，Link State ID和Advertising Router的组合共同标识一条LSA。

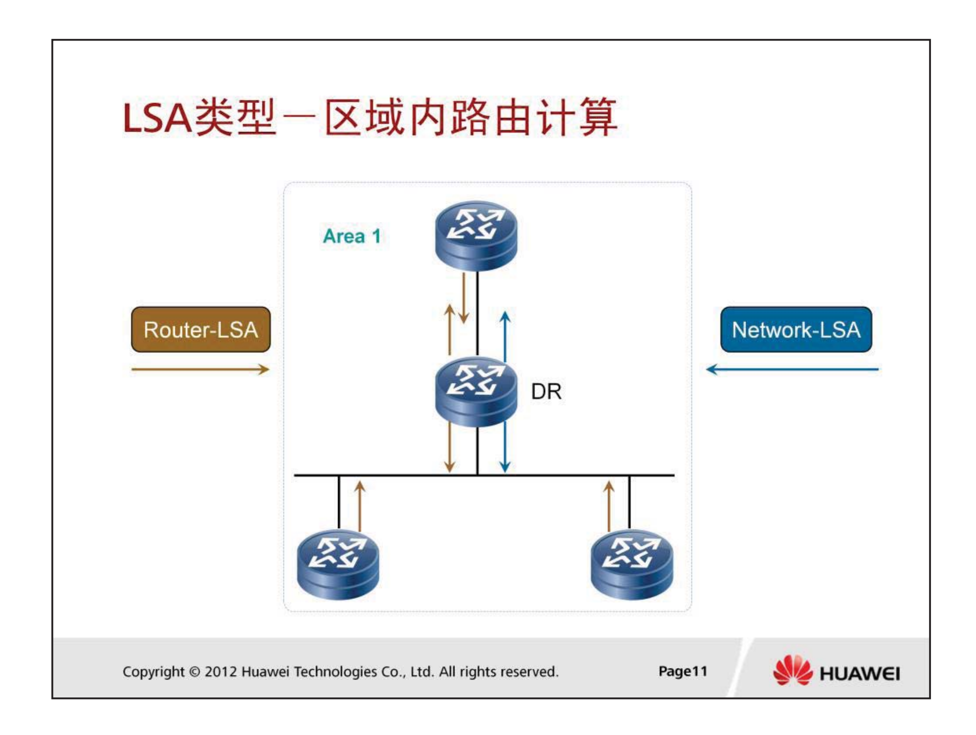
LSA类型－区域内路由计算

LS Type	LSA名称	LSA描述
1	Router-LSA	每一个路由器都会生成。这种LSA描述某区域内路由器端口链路状态的集合。只在所描述的区域内泛洪。
2	Network-LSA	由DR生成，用于描述广播型网络和NBMA网络。这种LSA包含了该网络上所连接路由器的列表。只在该网络所属的区域内泛洪。

 Page10



Router-LSA和Network-LSA用于计算区域内路由，这两类LSA描述的是具体的链路状态信息。



每台路由器都会向外发布Router-LSA。

只有DR向外发布Network-LSA。

LSA类型—区域间路由计算

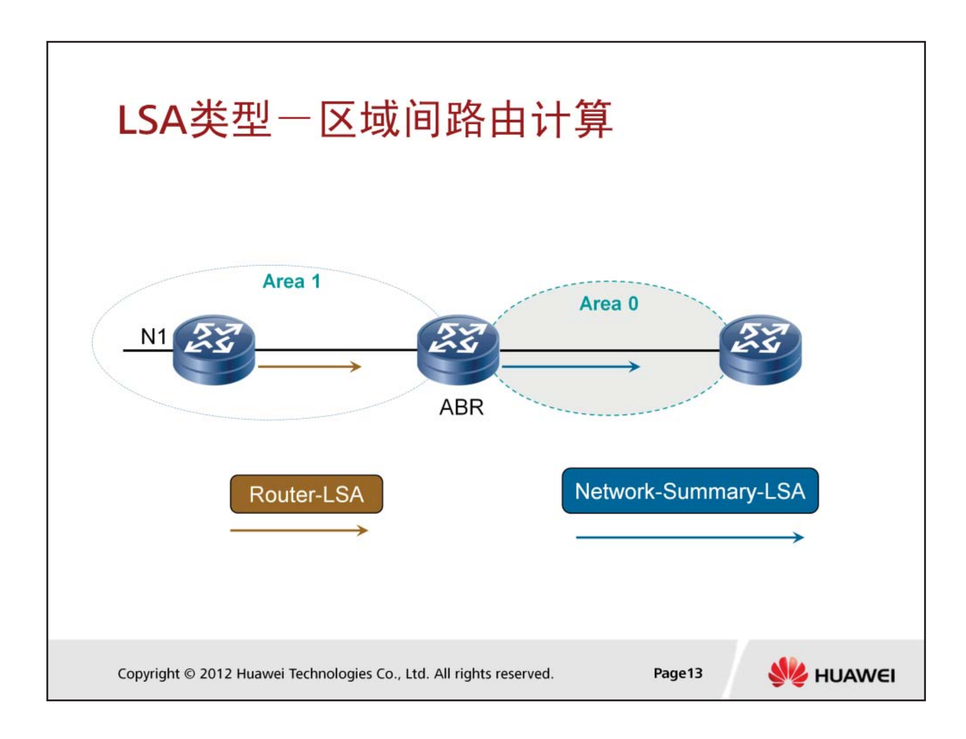
LS Type	LSA名称	LSA描述
3	Network-Summary-LSA	由区域边界路由器（ABR）产生，描述到AS内部本区域外部某一网段的路由信息，在该LSA所生成的区域内泛洪

Page12



Network-Summary-LSA用于计算区域间路由信息。这类LSA描述的是精简的路由信息，而不是详细的链路状态信息。

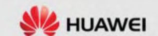
默认路由也可以通过Network-Summary-LSA发布。



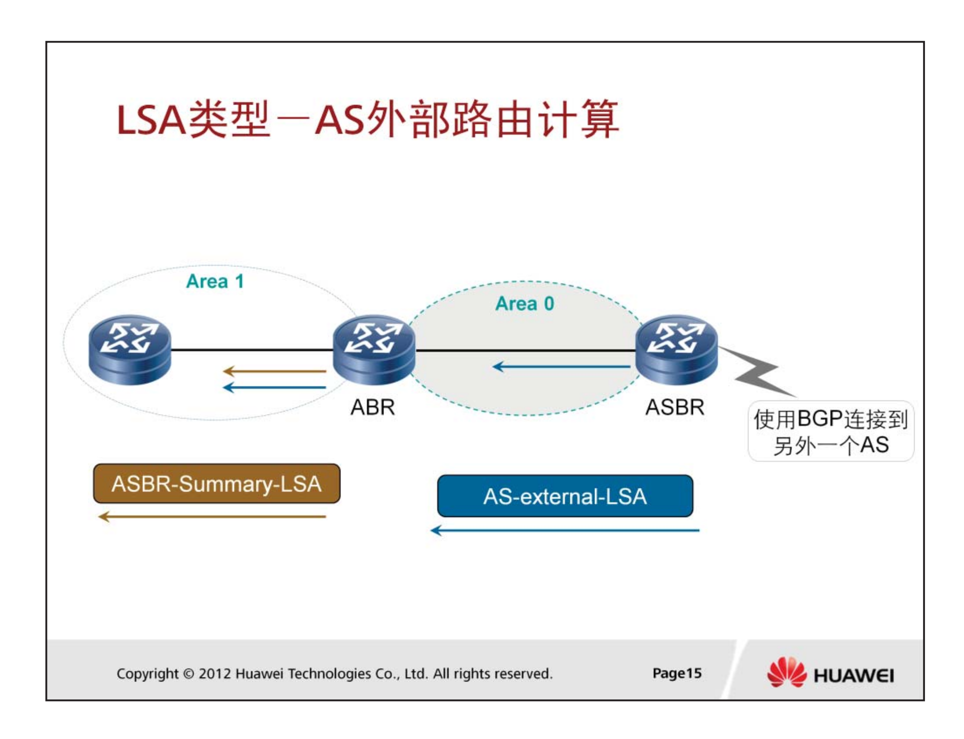
去往网段N1的路由通过Router-LSA发布到ABR，ABR将链路状态抽象成路由信息，通过Network-Summary-LSA发布到其它区域。

LSA类型－AS外部路由计算

LS Type	LSA名称	LSA描述
4	ASBR-Summary-LSA	由区域边界路由器（ABR）产生，描述到某一自治系统边界路由器（ASBR）的路由信息，在ABR所连接的区域内泛洪（ASBR所在区域除外）
5	AS-external-LSA	由自治系统边界路由器（ASBR）产生，描述到AS外部某一网段的路由信息，在整个AS内部泛洪

 Page14 

第四类用于描述如何到达ASBR，第五类由ASBR描述如何到达AS外部某网段，这两类LSA配合起来用于计算AS外部路由。



ASBR使用第五类LSA描述外部路由，这些第五类LSA在整个AS内部泛洪
。

当ABR向其它区域通告所接收到的第五类LSA时，同时为该区域生成一条第四类LSA描述如何到达ASBR。第四类LSA只能在一个区域内泛洪，第五类LSA每泛洪到一个区域，相关的ABR就要为该区域重新生成一条新的第四类LSA。

Link State ID

LSA名称	Link State ID
Router-LSA	生成这条LSA的路由器的Router ID
Network-LSA	所描述网段上DR的端口IP地址
Network-Summary-LSA	所描述的目的网段的地址
ASBR-Summary-LSA	所描述的ASBR的Router ID
AS-External-LSA	所描述的目的网段的地址

 Page16 

Link State ID是该LSA所描述链路的标识，对于不同类型的LSA，其含义也不同。

LS Sequence Number

```
[RTD]display ospf lsdb router self-originate

                OSPF Process 1 with Router ID 4.4.4.4
                         Link State Database

                Area: 0.0.0.2

Type      : Router
Ls id     : 4.4.4.4
Adv rtr   : 4.4.4.4
Ls age    : 1125
Len       : 48
Options   : (DC)
Seq#      : 80000008
Chksum    : 0x7b52
Link count: 2
   Link ID: 4.4.4.4
   Data   : 255.255.255.255
   Type   : StubNet
   Metric : 1
   Link ID: 10.1.3.1
   Data   : 10.1.3.2
   Type   : TransNet
   Metric : 1
```

序列号越大表示该LSA实例越新

 Page17 HUAWEI

LS type、Link State ID和Advertising Router的组合唯一标识一条LSA，但是对于一条LSA，有可能同时存在多个实例。LS sequence number用于检查哪一个实例更新。

LS Sequence Number:

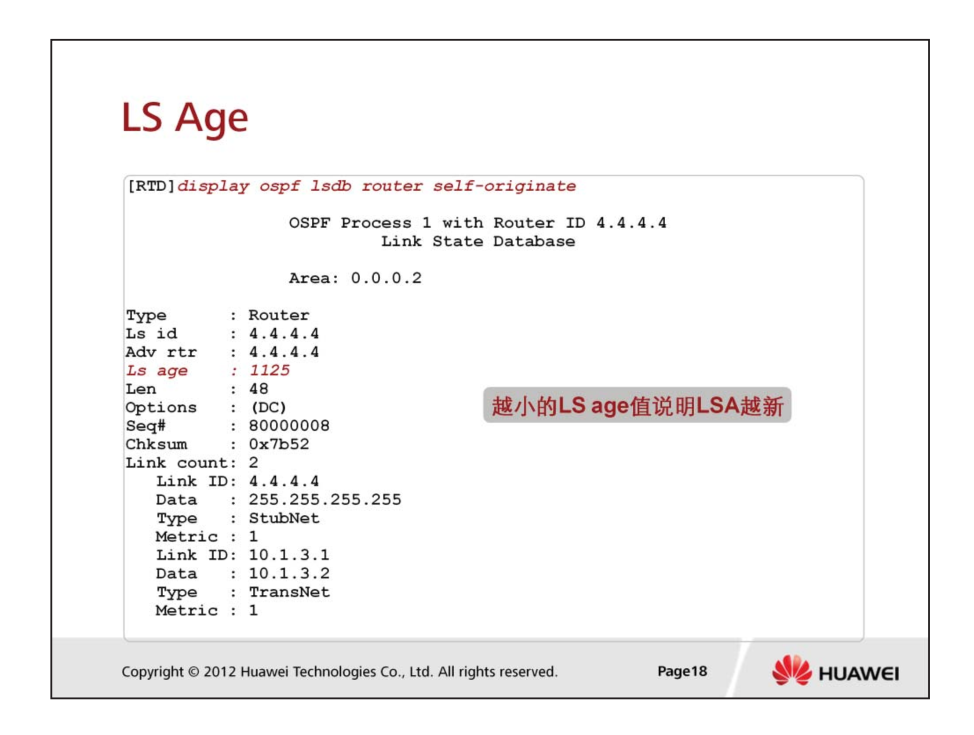
LS Sequence Number是一个32位的有符号整数，用于检测过期和重复的LSA。

由于LS Sequence Number是32位有符号整数，因此数值0x80000000，也就是-231是最小的数值，但此数值是被保留的，协议可用的最小数值为0x80000001（即-231＋1）。

当路由器生成一条新的LSA时，使用序列号0x80000001做为该LSA的初始序列号，此后，每次更新该LSA，序列号加1。

序列号越大表示该LSA实例越新。

当路由器收到一条自己产生的LSA，而且此LSA的LS Sequence Number比该路由器最近产生的这条LSA的LS Sequence Number更新时，路由器需要重新生成这条LSA的实例，其LS Sequence Number为收到的LSA中的LS Sequence Number加1。



LS Age：

此数值的单位是秒。在LSDB中的LSA的LS age随时间而增长。

一条LSA在向外泛洪之前，LS Age的值需要增加InfTransDelay（该值可以在端口上设置，缺省为1秒，表示在链路上传输的延迟）。

如果一条LSA的LS Age达到了LSRefreshTime（30分钟），这条LSA的生成者需要重新生成一个该LSA的实例，如果一条LSA的LS Age达到了MaxAge（1小时），这条LSA就要被删除。

LS Age数值越小表示此LSA越新。

如果路由器希望从网络中删除一条自己此前生成的LSA，则重新生成该条LSA的一个实例，将LS Age设置为Max Age即可。

如果路由器收到一条LS Age设置为Max Age的LSA，则从LSDB中删除此LSA（如果LSDB中存在此LSA）。



问 题

如何检查OSPF报文头是否合法?

如何配置OSPF报文的认证?

OSPF的LSA类型有哪些?

如何检测LSA的新旧?

 Page19



如何检查OSPF报文头是否合法?

检查版本号，Area ID，验证方法和验证信息。

如何配置OSPF报文的认证?

有区域验证和接口验证两种方式。

使用区域验证时，一个区域中所有路由器在该区域下的验证模式和口令必须一致；接口验证方式用于在相邻的路由器之间设置验证模式和口令，优先级高于区域验证方式。

OSPF的LSA类型有哪些?

基本的LSA有Router-LSA，Network-LSA，Network-Summary-LSA，ASBR-Summary-LSA和AS-External-LSA。

如何检测LSA的新旧?

使用LS Sequence Number和LS age，Sequence Number越大表示LSA越新，如果Sequence Number一致，则比较LS age，LS age越小表示LSA越新。


谢谢
www.huawei.com


建立OSPF邻居与邻接
关系
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍OSPF邻居与邻接关系的建立过程。

此过程也就是OSPF协议交互的过程，包括Hello报文，邻居状态变换以及链路状态数据库同步等内容。

Page1





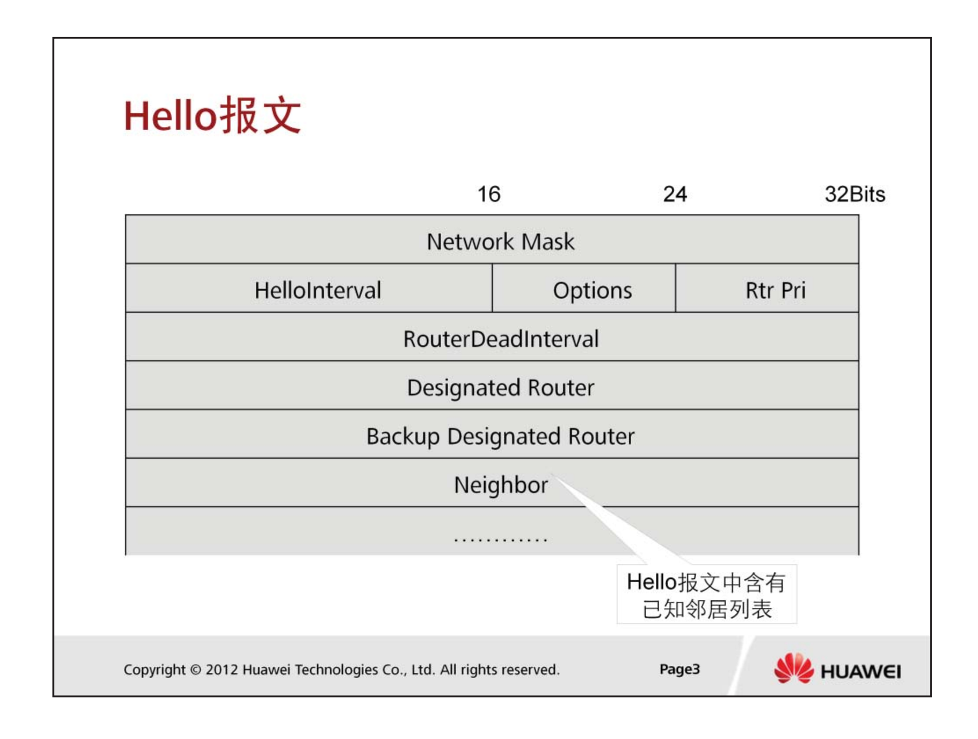
培训目标

学完本课程后，您应该能：

- 理解Hello报文的作用
- 理解OSPF邻居状态变换
- 理解邻居关系和邻接关系的建立过程
- 理解LSDB同步过程

Page2





重要字段解释：

Network Mask：发送Hello报文的接口的网络掩码。

HelloInterval：发送Hello报文的时间间隔。单位为秒。

Options：标识发送此报文的OSPF路由器所支持的可选功能。具体的可选功能不在本课程的讨论范围之列。

Rtr Pri：发送Hello报文的接口的Router Priority，用于选举DR和BDR。

RouterDeadInterval：宣告邻居路由器不继续在该网段上运行OSPF的时间间隔，单位为秒，通常为四倍HelloInterval。

Designated Router：发送Hello报文的路由器所选举出的DR的IP地址。如果设置为0.0.0.0，表示未选举DR路由器。

Backup Designated Router：发送Hello报文的路由器所选举出的BDR的IP地址。如果设置为0.0.0.0，表示未选举BDR路由器。

Neighbor：邻居路由器的Router ID列表。表示本路由器已经从该邻居收到合法的Hello报文。

验证合法Hello报文

16	24	32Bits
Network Mask		
HelloInterval	Options	Rtr Pri
RouterDeadInterval		
Designated Router		
Backup Designated Router		
Neighbor		
…………		

 Page4 HUAWEI

验证一个Hello报文是否合法之前首先需要验证一个OSPF报文是否合法。

验证一个接收到的Hello报文是否合法包括：

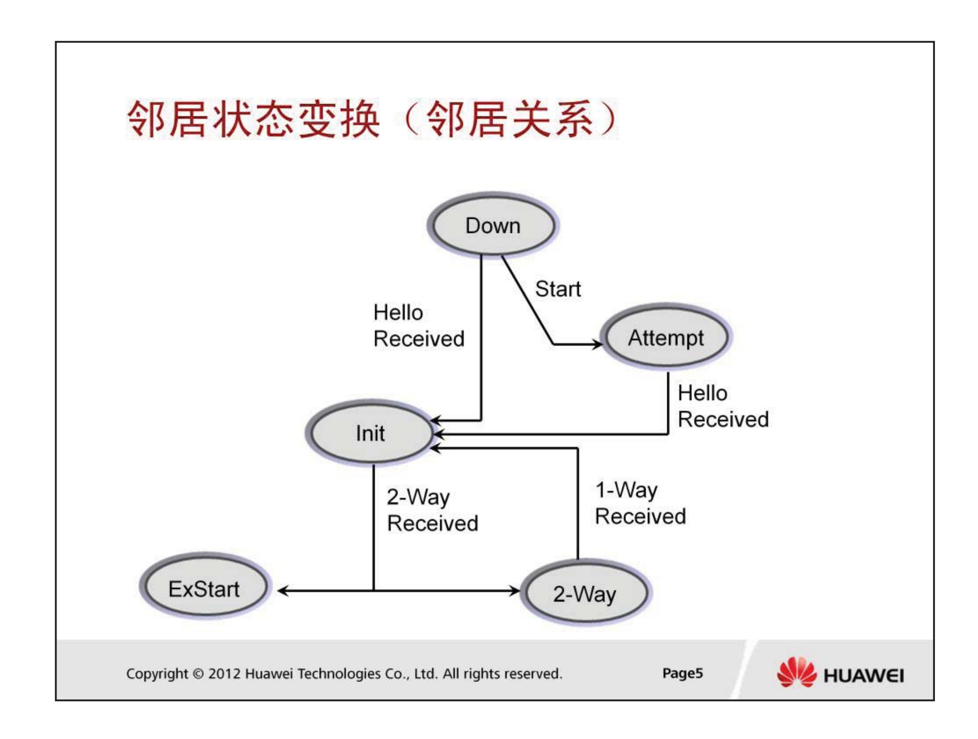
1. 如果接收端口的网络类型是广播型，点到多点或者NBMA，所接收的Hello报文中Network Mask字段必须和接收端口的网络掩码一致，如果接收端口的网络类型为点到点类型或者是虚连接，则不检查Network Mask字段；

2. 所接收的Hello报文中的HelloInterval字段必须和接收端口的配置保持一致；

3. 所接收的Hello报文中的RouterDeadInterval字段必须和接收端口的配置保持一致；

4. 所接收的Hello报文中的Options字段中的E-bit（表示是否接收外部路由信息）必须和相关区域的配置保持一致。关于此比特的具体意义将在《OSPF特殊区域》中详细解释。

如果路由器发现所接收的合法Hello报文的邻居列表中有自己的Router ID，则认为已经和邻居建立了双向连接，表示邻居关系已经建立。



这是形成邻居关系的过程和相关邻居状态的变换过程。

Down：这是邻居的初始状态，表示没有从邻居收到任何信息。在NBMA网络上，此状态下仍然可以向静态配置的邻居发送Hello报文，发送间隔为PollInterval，通常和RouterDeadInterval间隔相同。

Attempt：此状态只在NBMA网络上存在，表示没有收到邻居的任何信息，但是已经周期性的向邻居发送报文，发送间隔为HelloInterval。如果RouterDeadInterval间隔内未收到邻居的Hello报文，则转为Down状态。

Init：在此状态下，路由器已经从邻居收到了Hello报文，但是自己不在所收到的Hello报文的邻居列表中，表示尚未与邻居建立双向通信关系。在此状态下的邻居要被包含在自己所发送的Hello报文的邻居列表中。

2-WayReceived：此事件表示路由器发现与邻居的双向通信已经开始（发现自己在邻居发送的Hello报文的邻居列表中）。Init状态下产生此事件之后，如果需要和邻居建立邻接关系则进入ExStart状态，开始数据库同步过程，如果不能与邻居建立邻接关系则进入2-Way。

2-Way：在此状态下，双向通信已经建立，但是没有与邻居建立邻接关系。这是建立邻接关系以前的最高级状态。

1-WayReceived：此事件表示路由器发现自己没有在邻居发送Hello报文的邻居列表中，通常是由于对端邻居重启造成的。

查看邻居状态

```
[RTD]display ospf peer

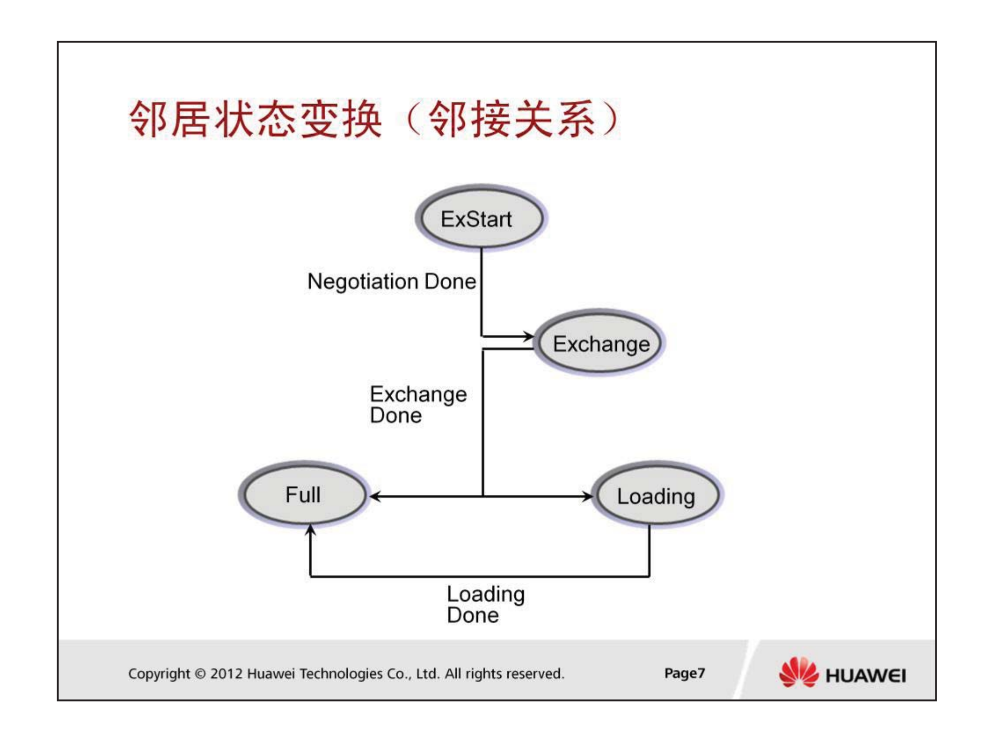
                OSPF Process 1 with Router ID 4.4.4.4
                               Neighbors

 Area 0.0.0.1 interface 10.1.1.4(Ethernet0/0)'s neighbor(s)
 RouterID: 1.1.1.1         Address: 10.1.1.1
       State: 2 Way  Mode: None    Priority: 1
       DR: 10.1.1.2  BDR: 10.1.1.3
       Dead timer due in 40  sec
       Retrans timer interval: 4
       Neighbor is up for 01:06:37
       Authentication Sequence: [ 0 ]
```


Page6

HUAWEI

本例中，网络中的DR为10.1.1.2，BDR为10.1.1.3，RTD和1.1.1.1都是DROther，所以RTD不能和1.1.1.1建立邻接关系。在不能建立邻接关系的情况下，邻居稳定状态为2 Way。



DD 序列号（DD Sequence Number）：

每一个DD报文都有一个DD序列号，用于DD报文的确认机制。DD序列号是一个两字节的值。

主从关系（Master/Slave）：

当两个路由器之间通过DD报文交换数据库信息的时候，首先形成一个主从关系，Router ID大的优先为主，确认主从关系之后，主路由器发送DD报文，从路由器不能主动发送DD报文，只能回应主路由器发送的DD报文，回应时使用的DD序列号必须和所回应的主路由器发送的DD报文的序列号一致。

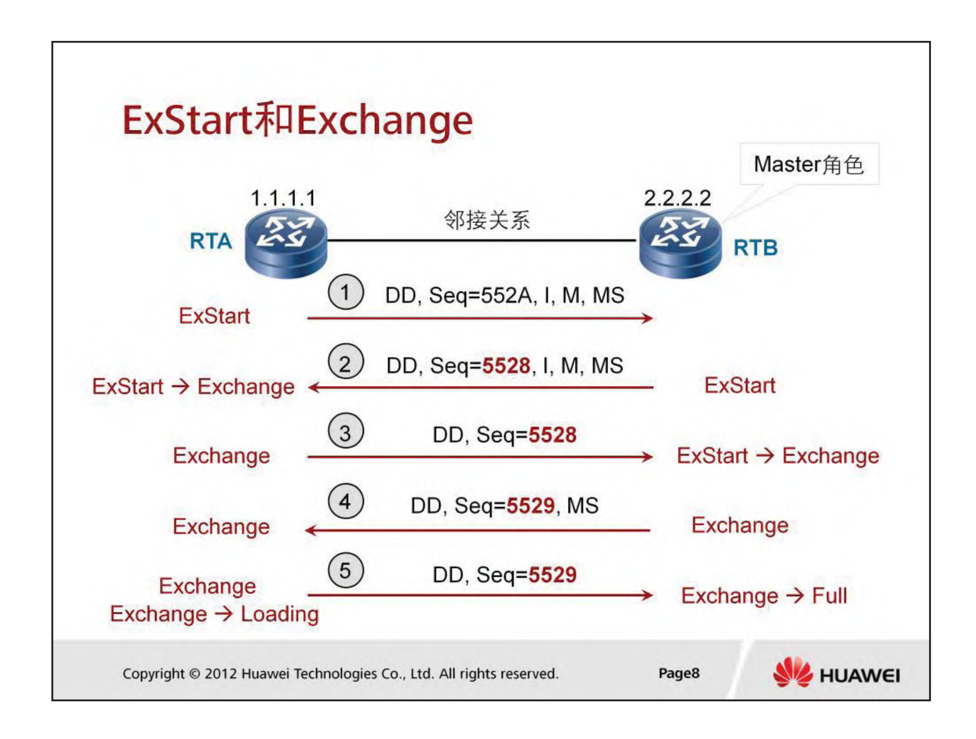
相关状态解释如下：

ExStart：这是形成邻接关系的第一个步骤，邻居状态变成此状态以后，路由器开始向邻居发送DD报文。主从关系是在此状态下形成的；初始DD序列号是在此状态下决定的。在此状态下发送的DD报文不包含链路状态描述。

Exchange：此状态下路由器相互发送包含链路状态信息摘要的DD报文，描述本地LSDB的内容。

Loading：相互发送LS Request报文请求LSA，发送LS Update通告LSA。

Full：两路由器的LSDB已经同步。



1. 邻居状态机变为ExStart以后，RTA向RTB发送第一个DD报文，在这个报文中，DD序列号被设置为552A（假设），Initial比特为1表示这是第一个DD报文，More比特为1表示后续还有DD报文要发送，Master比特为1表示RTA宣告自己为主路由器。

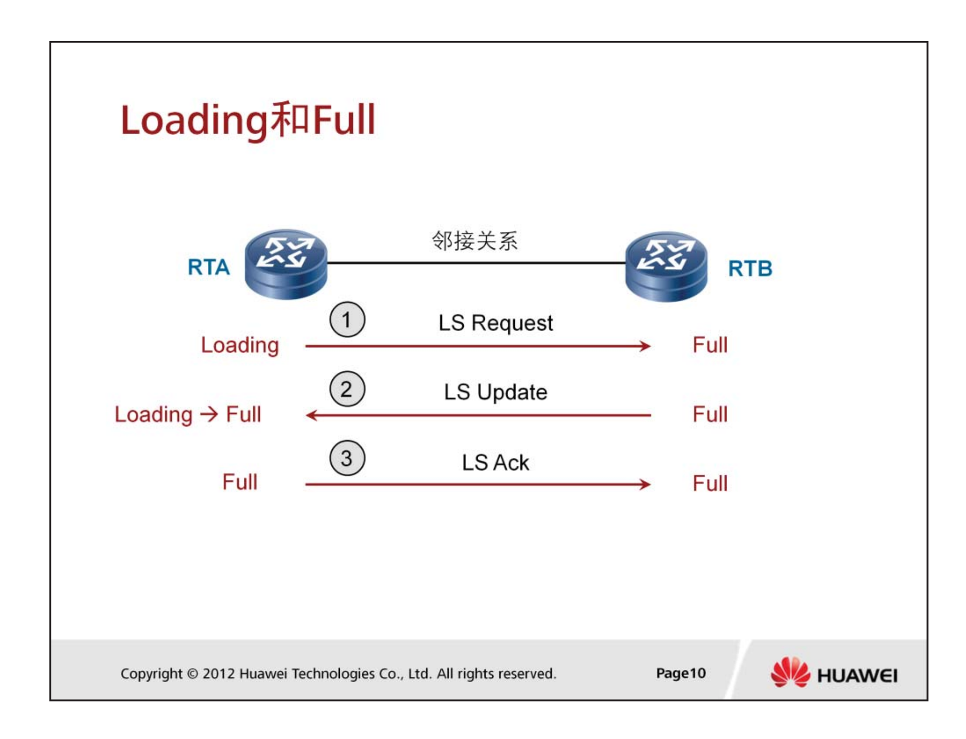
2. 邻居状态机变为ExStart以后，RTB向RTA发送第一个DD报文，在这个报文中，DD序列号被设置为5528（假设）。由于RTB的Router ID比RTA的大，所以RTB应当为主路由器，Router ID的比较结束后，RTA会产生一个NegotiationDone的事件，所以RTA将状态机从ExStart改变为Exchange。

3. 邻居状态机变为Exchange以后，RTA发送一个新的DD报文，在这个新的报文中包含LSDB的摘要信息，序列号设置为RTB在步骤2里使用的序列号，More比特为0表示不需要另外的DD报文描述LSDB，Master比特为0表示RTA宣告自己为从路由器。收到这样一个报文以后，RTB会产生一个NegotiationDone的事件，因此RTB将邻居状态改变为Exchange。

4. 邻居状态变为Exchange以后，RTB发送一个新的DD报文，该报文中包含LSDB的描述信息，DD序列号设为5529（上次使用的序列号加1）。

5. 即使RTA不需要新的DD报文描述自己的LSDB，但是做为从路由器，RTA需要对主路由器RTB发送的每一个DD报文进行确认。所以，RTA向RTB发送一个新的DD报文，序列号为5529，该报文内容为空。

发送完最后一个DD报文之后，RTA产生一个ExchangeDone事件，将邻居状态改变为Loading；RTB收到最后一个DD报文之后，改变状态为Full（假设RTB的LSDB是最新最全的，不需要向RTA请求更新）。



1. 邻居状态变为Loading之后，RTA开始向RTB发送LS request报文，请求那些在Exchange状态下通过DD报文发现的，而且在本地LSDB中没有的链路状态信息。

2. RTB收到LS Request报文之后，向RTA发送LS Update报文，在LS Update报文中，包含了那些被请求的链路状态的详细信息。RTA收到LS Update报文之后，将邻居状态从Loading改变成Full。

3. RTA向RTB发送LS Ack报文，确保信息传输的可靠性。

LS Ack报文用于泛洪对已接收LSA的确认。

邻居状态变成Full，表示达到完全邻接状态。

查看OSPF邻居状态

```
[RTA]display ospf peer

          OSPF Process 1 with Router ID 1.1.1.1
                          Neighbors
 Area 0.0.0.0 interface 10.1.1.1(Ethernet0/0)'s neighbor(s)
 RouterID: 2.2.2.2          Address: 10.1.1.2
       State: Full  Mode: Nbr is Master  Priority: 1
       DR: 10.1.1.1  BDR: 10.1.1.2
       Retrans timer interval: 4
       Dead timer expires in 35s
       Neighbor has been up for 04:35:02
       Authentication Sequence: [ 0 ]
```


Page11 HUAWEI

RouterID：邻居的Router ID；

Address：邻居在该网段上的IP地址；

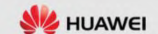
State：邻居状态；如果状态为Full，表示邻接关系完全建立。

Mode：交换DD报文时，邻居角色是Master还是Slave。

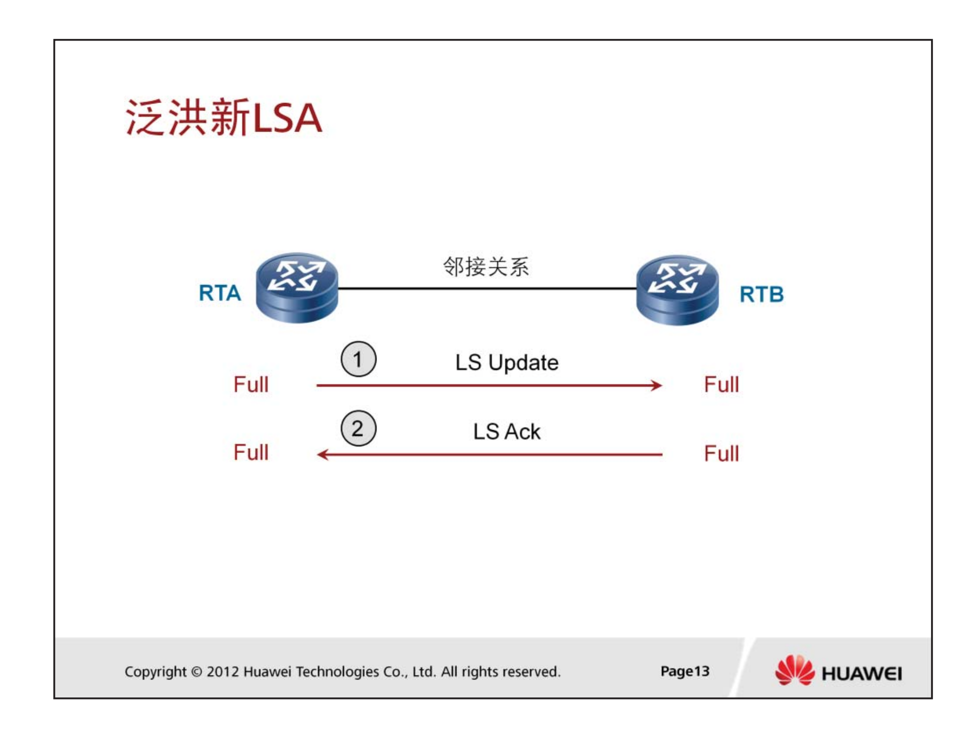
包含在各种报文中的LSA信息

Packet类型	LSA信息
Database Description	LSA头部信息，包括LS Type, LS ID, Advertising Router和LS Sequence Number
LS Request	只有LS Type, LS ID和Advertising Router
LS Update	完整的LSA信息，包括LSA头部和具体的链路状态信息
LS Ack	LSA头部信息，包括LS Type, LS ID, Advertising Router和LS Sequence Number

 Page12



不同的协议报文中包含不同部分的LSA信息。



当有新的LSA生成或收到时，这条新的LSA应当被泛洪。

泛洪新的LSA时，只需要使用LS Update报文和LS Ack报文。

1. 当RTA有新的LSA要泛洪时，RTA向RTB发送一个LS Update报文，在这个报文里包含这条LSA。

2. 收到新的LSA以后，RTB向RTA泛洪一个LS Ack报文进行确认。

当在两个处于完全邻接状态（邻居状态为Full）的路由器之间泛洪新的LSA时，邻居状态不受影响。

OSPF报文的目的地址

	Hello	Database Description	Link State Request	Link State Update	Link State Ack
Point-to-point	224.0.0.5	224.0.0.5	224.0.0.5	224.0.0.5	224.0.0.5
NBMA	单播	单播	单播	单播	单播
Virtual link	单播	单播	单播	单播	单播

 Page14



点到点网段上所有报文发送到组播地址224.0.0.5（AllSPFRouters）。

NBMA网段上所有报文以单播形式发送，目的地是已经手工配置好的邻居。

虚连接的报文以单播形式发送。

OSPF报文的目的地址

	Hello	Database Description	Link State Request	Link State Update	Link State Ack
Broadcast	224.0.0.5	单播	单播	224.0.0.5 或 224.0.0.6	224.0.0.5 或 224.0.0.6
Point-to-MultiPoint	224.0.0.5	单播	单播	224.0.0.5 或 单播	224.0.0.5

Page15


广播型网段上，DR和BDR发送LS Update报文和LS Ack报文的目的地址是224.0.0.5（AllSPFRouters），其余路由器发送LS Update报文和LS Ack报文的目的地址是224.0.0.6（AllDRouters）。

点到多点网段上，如果LS Update报文是对LS Request报文的回应，则该LS Update报文以单播形式发送给邻居；如果发送LS Update报文是为了泛洪新的LSA，则该LS Update报文的目的地址为224.0.0.5（AllSPFRouters）。



问 题

如何验证一个Hello报文是否合法?

邻居状态变换分为几个阶段?

不能建立邻接关系的情况下，邻居稳定工作状态是什么?

可以建立邻接关系的情况下，邻居稳定工作状态是什么?

 Page16



如何验证一个Hello报文是否合法?

检查Network Mask，HelloInterval，RouterDeadInterval以及Options字段中的E-bit。

邻居状态变换分为几个阶段?

分为两个阶段，第一阶段建立邻居关系，第二阶段建立邻接关系。

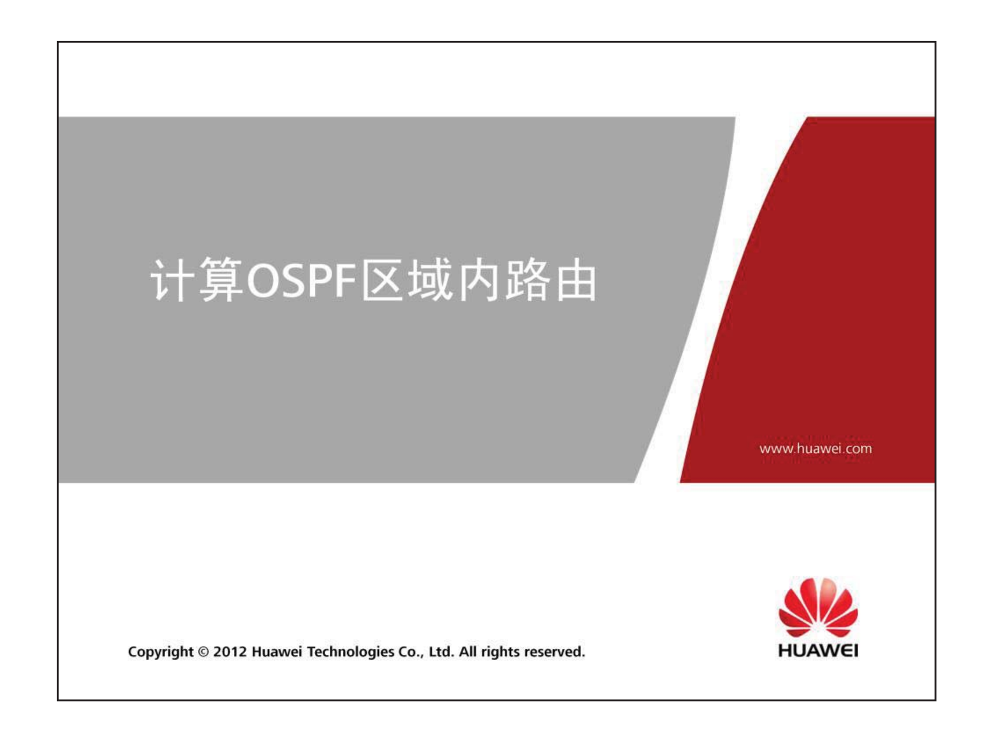
不能建立邻接关系的情况下，邻居稳定工作状态是什么?

不能建立邻接关系的情况下，邻居稳定工作状态是2 way。

可以建立邻接关系的情况下，邻居稳定工作状态是什么?

可以建立邻接关系的情况下，邻居稳定工作状态是Full，表示邻接关系已经完全建立。


谢谢
www.huawei.com


计算OSPF区域内路由
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍OSPF如何计算区域内路由。

课程内容包括如何使用Router-LSA和Network-LSA表示链路状态信息，以及如何计算最短路径树等。

Page1



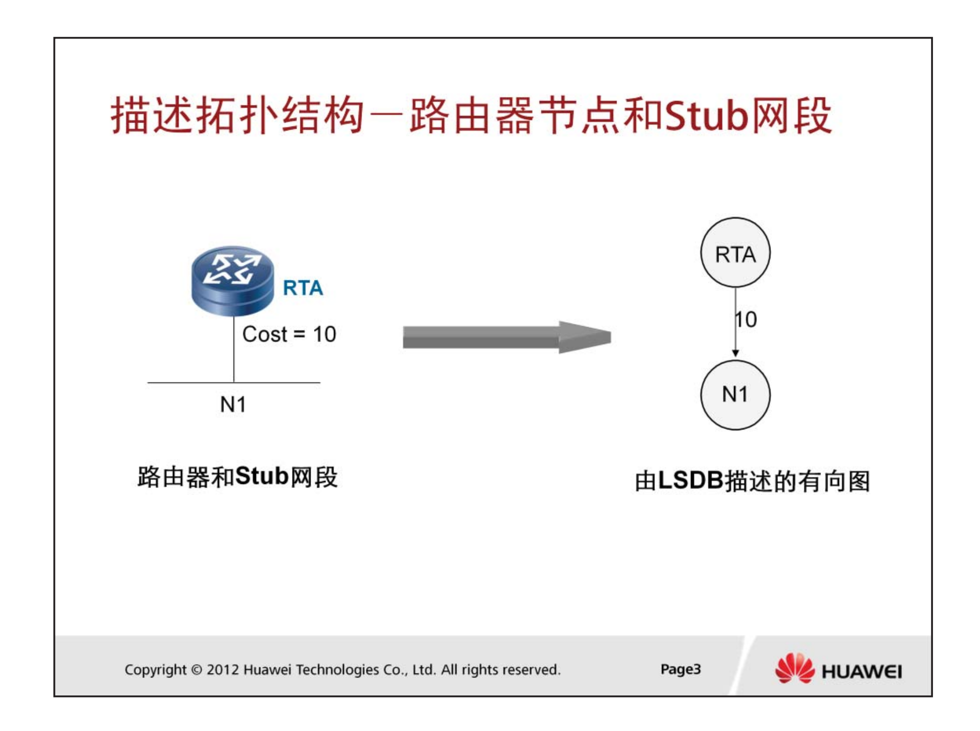


学完本课程后，您应该能：

- 理解Router-LSA
- 理解Network-LSA
- 理解最短路径树的计算

 Page2



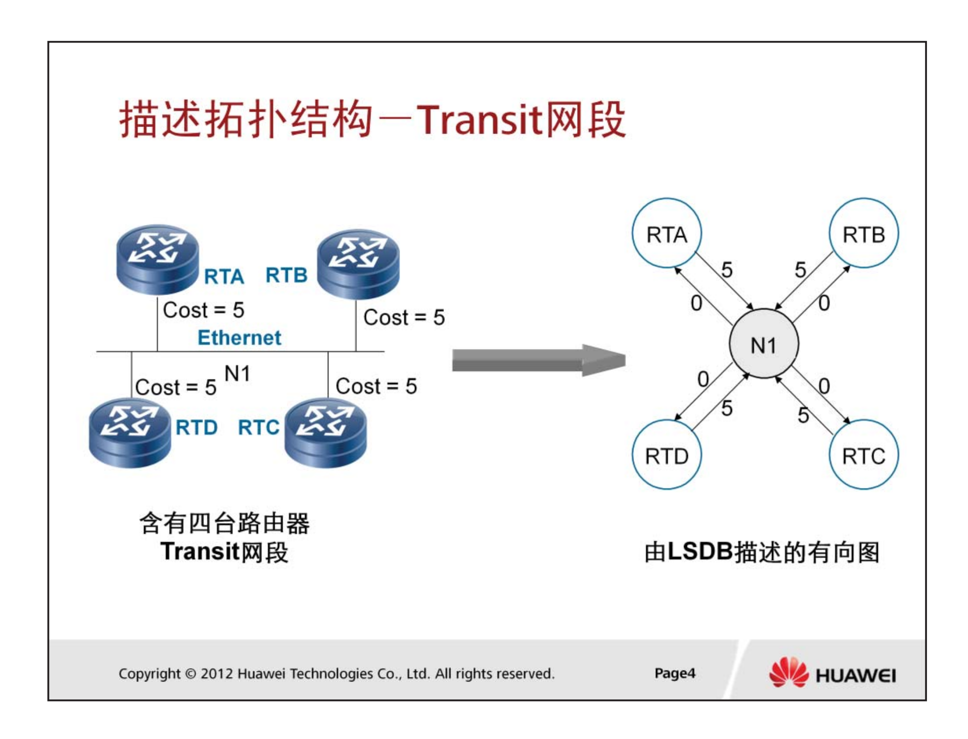


LSDB通过描述一个有向线段图来描述网络拓扑结构，该有向图的端点有三种类型：路由器节点，Stub网段和Transit网段。

Stub网段表示该网段只有数据入口，例如一个Loopback接口就是一个Stub网段。

此胶片描述了路由器节点和Stub网段的表示方式。

Cost表示从一个端点到另一个端点的开销，该参数可以在OSPF接口上配置，表示数据离开该接口（出接口）的开销。

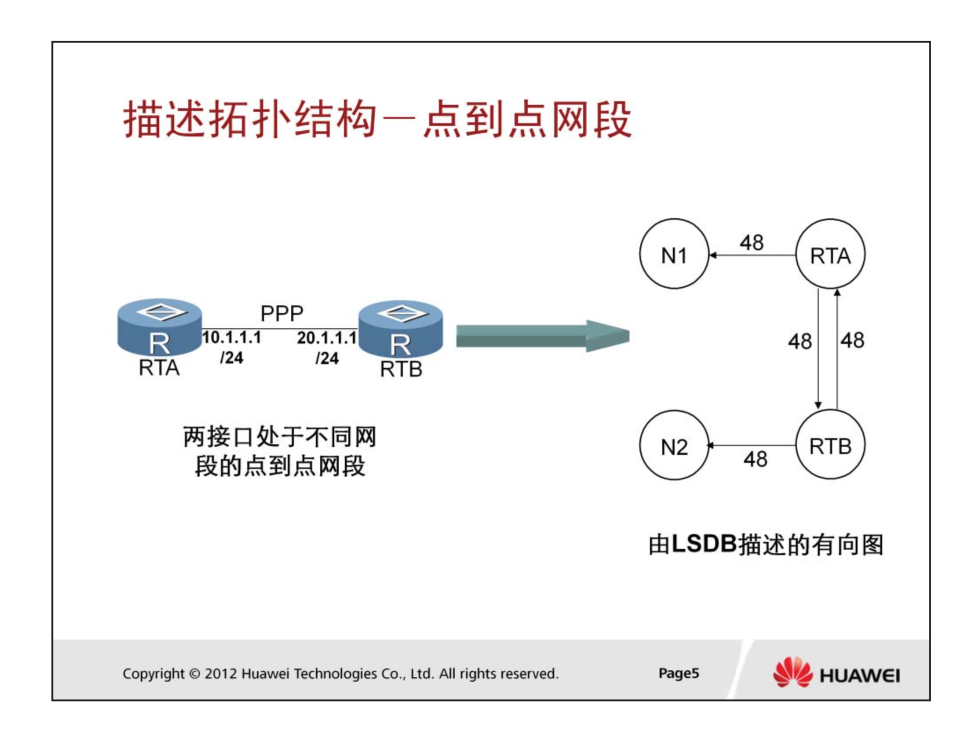


Transit网段有能力转发既不是本网段产生的，也不以本网段做为目的地的数据。

有至少两台路由器的广播型网段或NBMA网段就是一种Transit网段。

从路由器到所连Transit网段的开销值就是连接到这个网段的接口所配置的开销值。

从一个Transit网段到连接到这个网段的路由器的开销为0。



本例中：

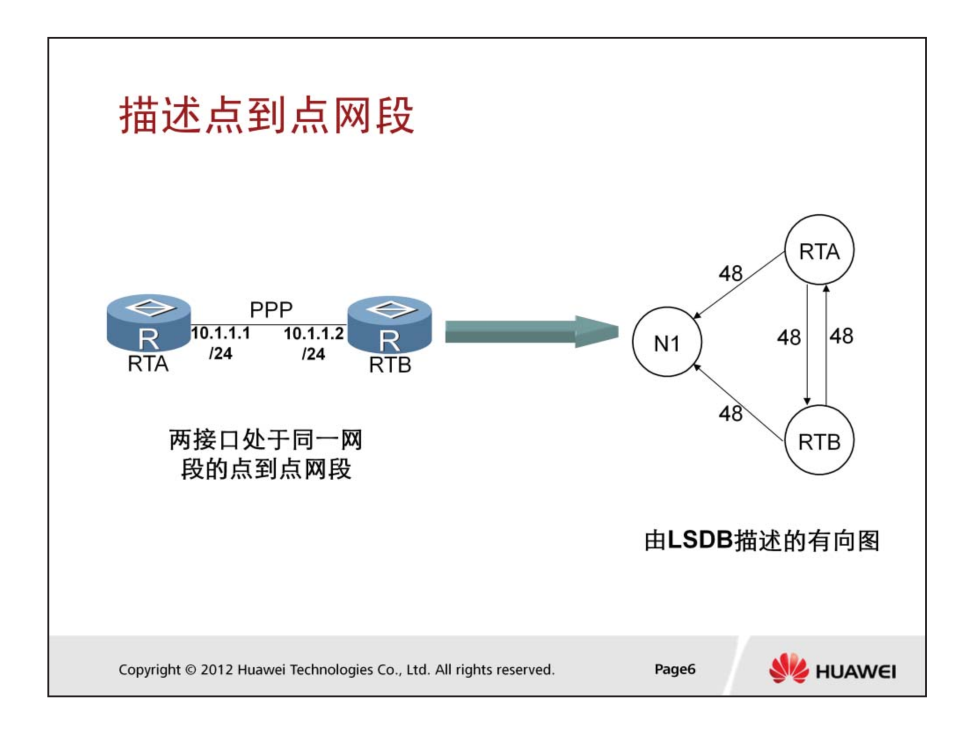
两接口的开销配置为48。有向图中N1表示10.1.1.0/24；N2表示20.1.1.0/24。

LSDB描述两接口处于不同网段的点到点网段的规则如下：

两台路由器经由两条有向线段直接相连，每个方向一条。

两个接口的网段被表示成Stub网段。

每个路由器通告一个Stub连接到该路由器所连的网段。



本例中:

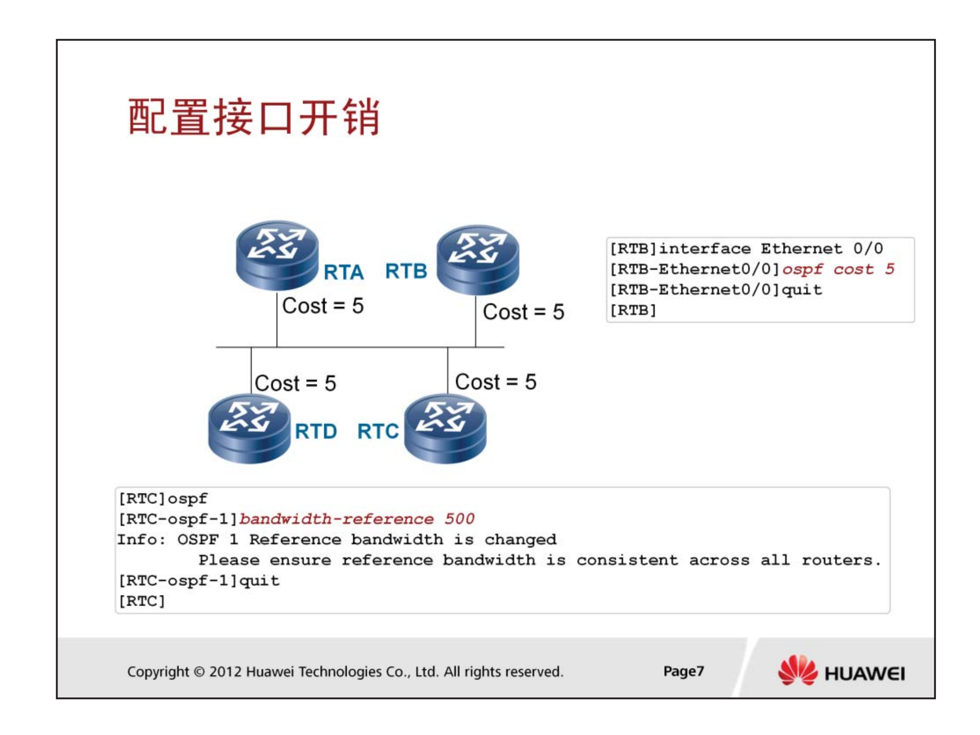
两接口的开销配置为48。有向图中N1表示10.1.1.0/24。

LSDB描述两接口处于同一网段的点到点网段的规则如下:

两台路由器经由两条有向线段直接相连，每个方向一条。

连接两个接口的网段被表示成Stub网段。

两个路由器同时通告Stub连接到该PPP网段。



默认情况下，OSPF接口开销与接口的带宽有关，计算公式为：bandwidth-reference / bandwidth。

bandwidth-reference默认取值为100M，bandwidth（带宽）的单位使用bit/s。因此，一个FE接口的开销默认为1。

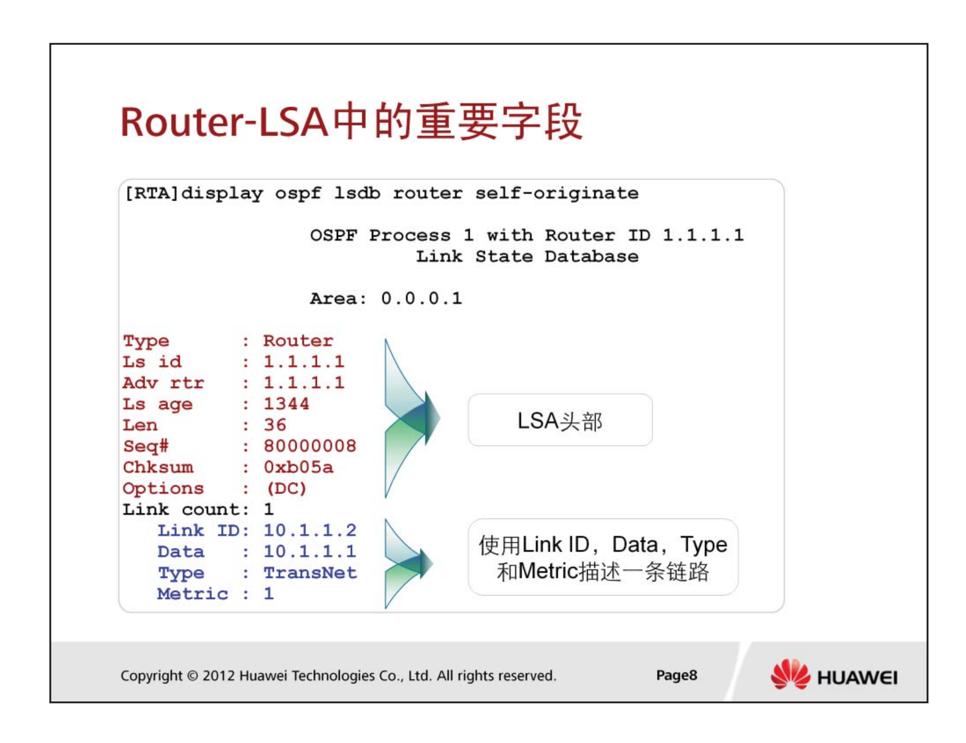
可以通过两种方式修改接口开销，第一种方式是在接口模式下通过ospf cost命令直接修改；第二种方法是在OSPF进程模式下修改bandwidth-reference，由系统自动计算所有接口的新的开销值。

ospf cost cost

cost：OSPF接口的开销值，取值范围为1～65535。

bandwidth-reference value

value：计算OSPF接口开销时所依据的参考值。单位Mbit/s，取值范围1～2147483648。



每台OSPF路由器只使用一条Router-LSA描述属于一个区域的本地活动链接状态，一条Router-LSA可以描述多条链接，每条链接由Link ID，Data，Type和Metric描述。

1. Type：链接类型（并非OSPF所支持的网络类型），Router-LSA描述的链接类型共有四种：

Point-to-Point：描述一个从本路由器到邻居路由器之间的点到点链接。

TransNet：描述一个从本路由器到一个Transit网段（例如广播型网段或者NBMA网段）的链接。

StubNet：描述一个从本路由器到一个Stub网段（例如Loopback接口）的链接。

Virtual：表示这是一个从本路由器到虚连接对端ABR的链接。

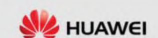
2. Link ID：此链接的对端标识，不同链接类型的Link ID表示的意义也不同。

3. Data：用于描述此链接的附加信息，不同的链接类型所描述的信息也不同。

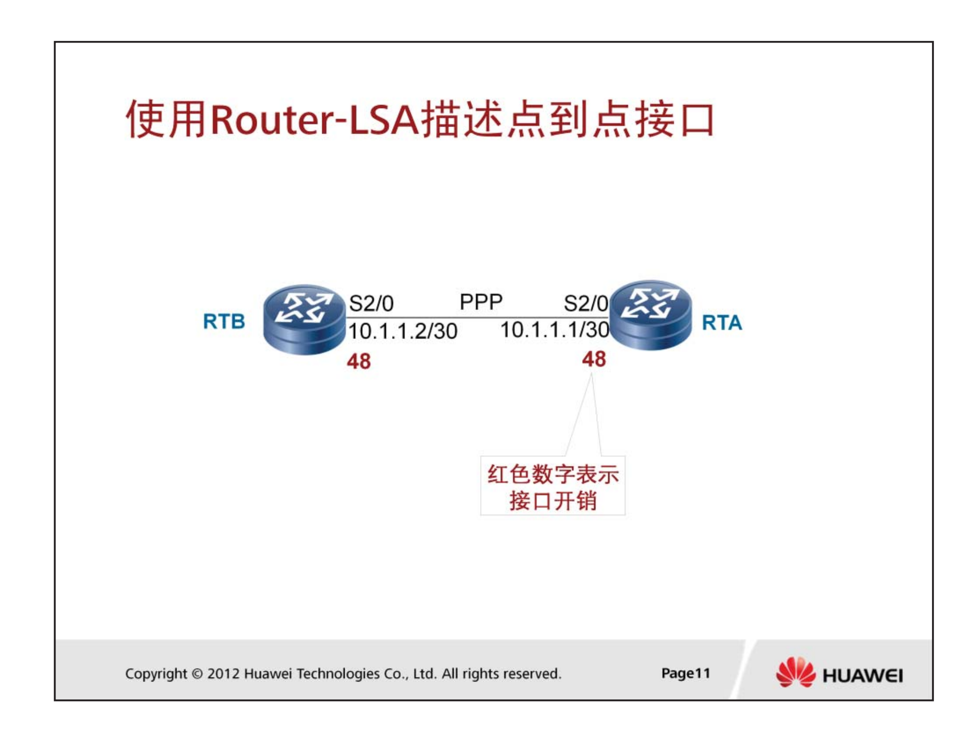
4. Metric：描述此链接的开销。

Router-LSA中的重要字段

Type	Link ID	Data
Point-to-point	邻居的Router ID	该网段上本地接口的IP地址
TransNet	DR的接口IP地址	该网段上本地接口的IP地址
StubNet	该Stub网段的IP网络地址	该Stub网段的网络掩码
Virtual	虚连接邻居的Router ID	去往该虚连接邻居的本地接口的IP地址

 Page10 

对于不同的链接类型，Link ID和Data所描述的内容也不同，该表格列出了相关的对应关系。



本例中:

RTA的Router ID是1.1.1.1；RTB的Router ID是2.2.2.2。

红色数字标注的是接口开销。

两个接口在同一个IP网段。

将该网段配置在Area 1。

使用Router-LSA描述点到点接口

```
[RTA]dis ospf lsdb router self-originate

         OSPF Process 1 with Router ID 1.1.1.1
                         Area: 0.0.0.1
                 Link State Database
    Type       : Router
    Ls id      : 1.1.1.1
    Adv rtr    : 1.1.1.1
    Ls age     : 12
    Len        : 48
    Options    :  E
    seq#       : 80000002
    chksum     : 0x5864
    Link count: 2
       Link ID: 2.2.2.2
       Data    : 10.1.1.1
       Link Type: P-2-P
       Metric : 48
       Link ID: 10.1.1.0
       Data    : 255.255.255.252
       Link Type: StubNet
       Metric : 48
```

去往该邻居的点到点连接

去往该点到点网段的Stub连接

Page12 HUAWEI

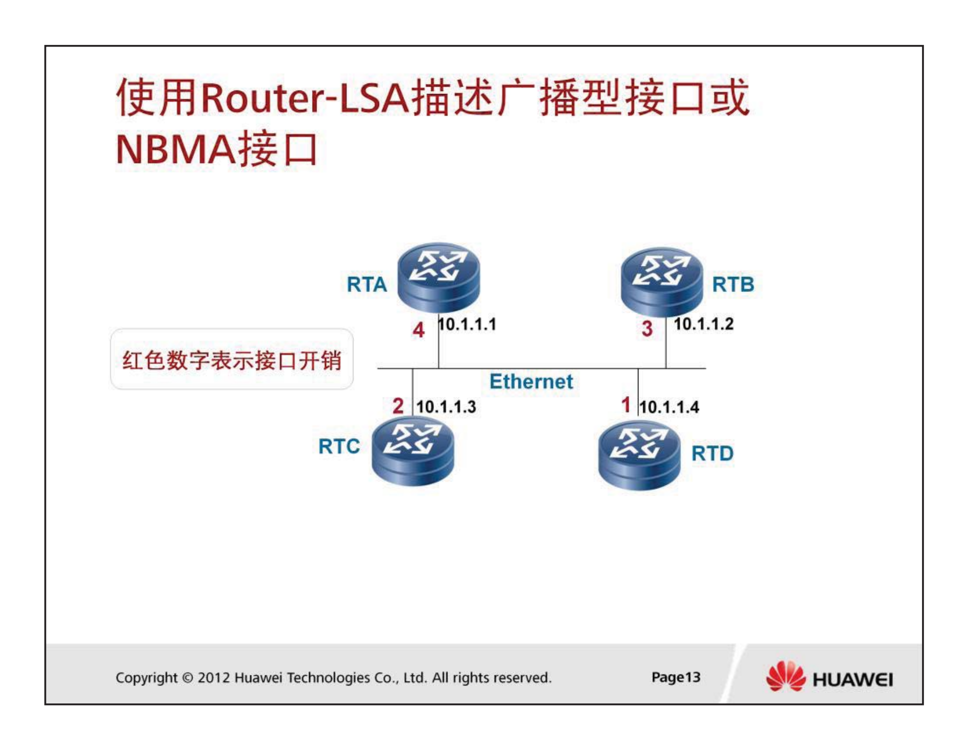
这是由RTA产生的Router-LSA。

在描述点到点接口的Router-LSA中：

1. 通告一个到邻居路由器的点到点链接，Link ID设置为对端的Router ID，Data设置为本地接口的IP地址；

2. 通告一个到该点到点网段的Stub连接，Link ID设置为该点到点网段的网络号，Data设置为该点到点网段的网络掩码；

3. 上述两个连接的Cost值均为该点到点接口上的Cost值。



本例中:

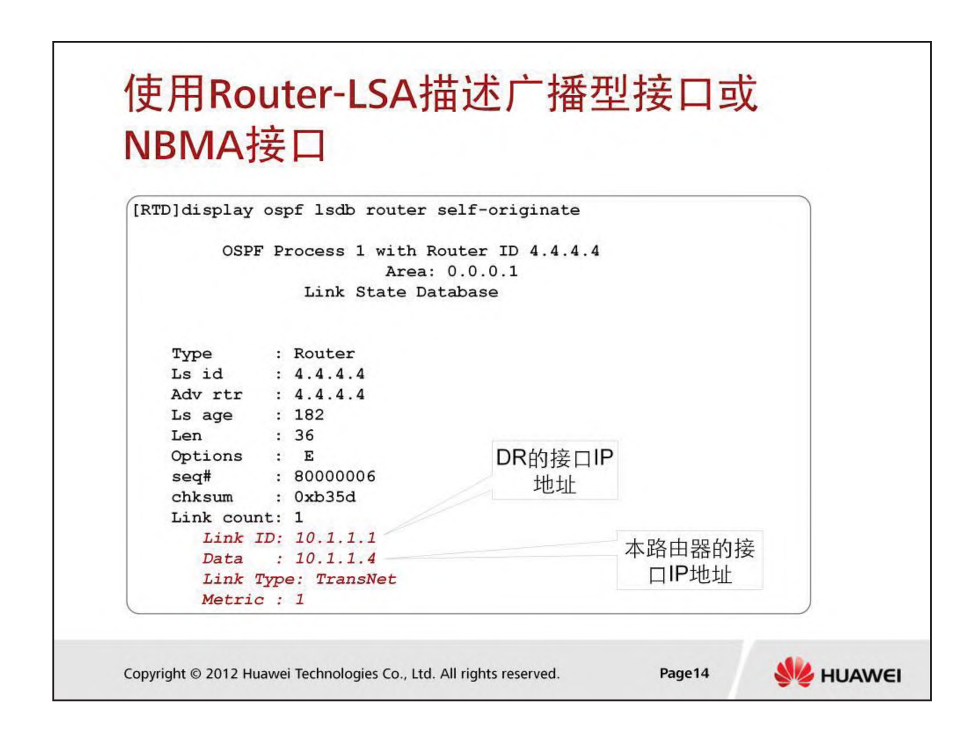
RTA的Router ID是1.1.1.1，RTB是2.2.2.2，RTC是3.3.3.3，RTD是4.4.4.4
。

红色数字表示接口开销。

通过配置接口的Router Priority将RTA配置为DR。

四个接口在同一个IP网段。

将该网段配置在Area 1中。



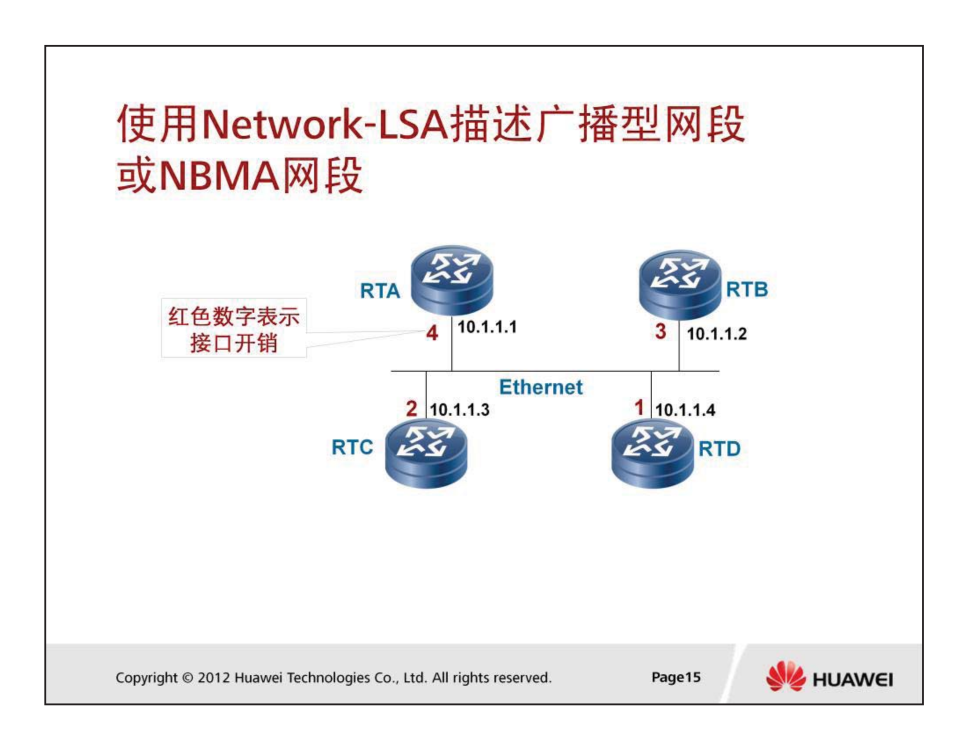
这是由RTD产生的Router-LSA。

在描述广播型或NBMA型接口的Router-LSA中：

1. 如果接口状态是Waiting，或者该网段上只有一个运行OSPF的路由器，或者该网段上没有DR，则通告一个通往该网段的Stub链接，Link ID设置为该网段的IP网络号，Link Data设置为该网段的网络掩码；

其他情况下，通告一个通往该网段的Transit连接，Link ID设置为DR的接口IP地址，Link Data设置为本地接口的IP地址。

2. 连接的开销值为接口的开销。

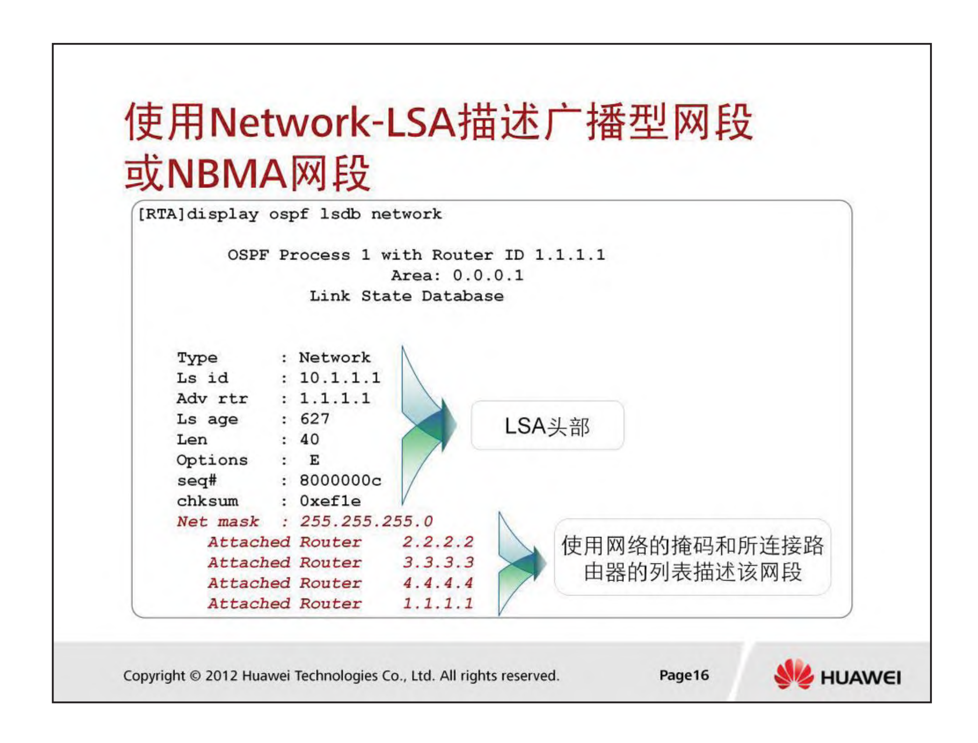


本例中：

RTA的Router ID为1.1.1.1，RTB为2.2.2.2，RTC为3.3.3.3，RTD为4.4.4.4
。

红色数字表示接口开销。

将该网段配置在Area 1中。



这是在Area 1中产生的Network-LSA。

在描述广播型网段或者NBMA网段的Network-LSA中：

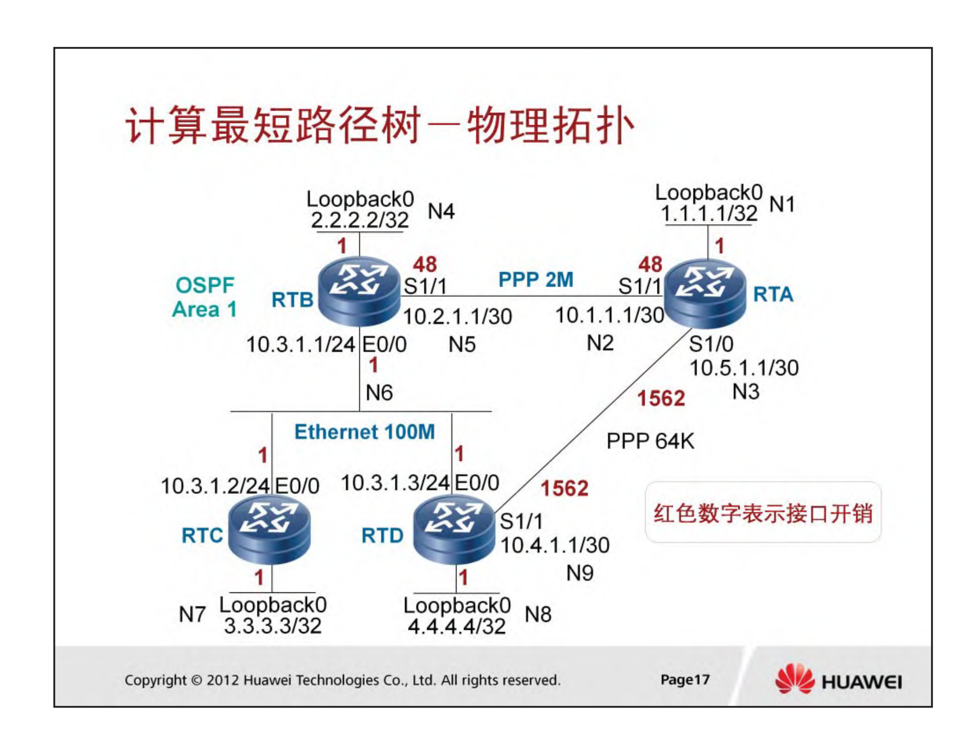
Link State ID设置为DR的接口IP地址。

Net mask设置为该网段的网络掩码。

Link State ID和Net mask做与运算，即可得出该网段的IP网络号。

在该LSA中，还包含一个连接到该网段的路由器列表。

从一个Transit网段到所连接的路由器的连接没有开销。



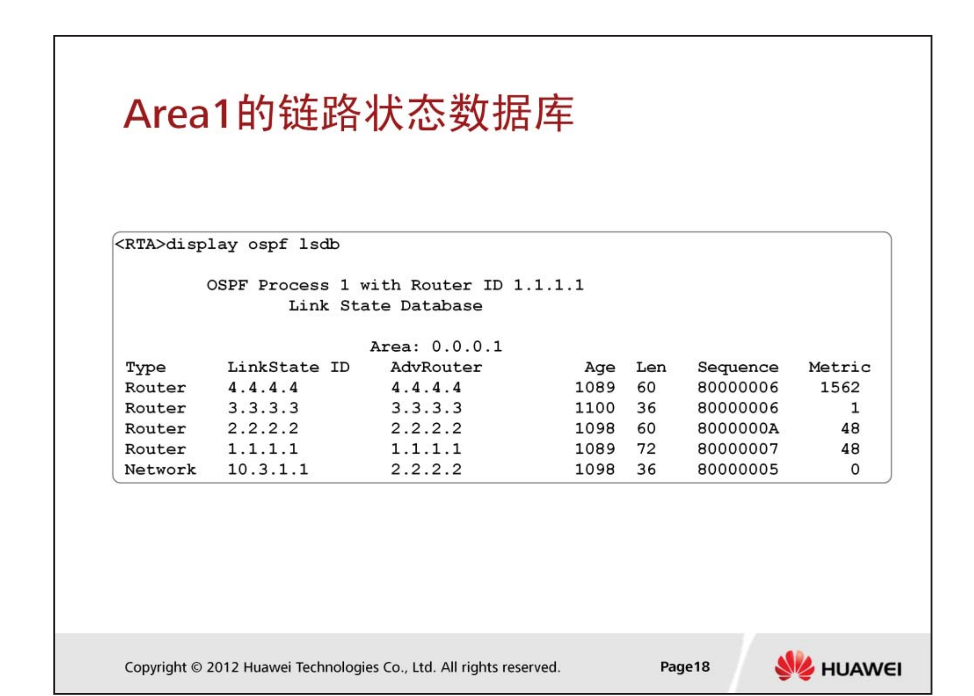
在本例中：

所有网段均在Area 1中。

网络中共有两个点到点网段和一个广播型网段。

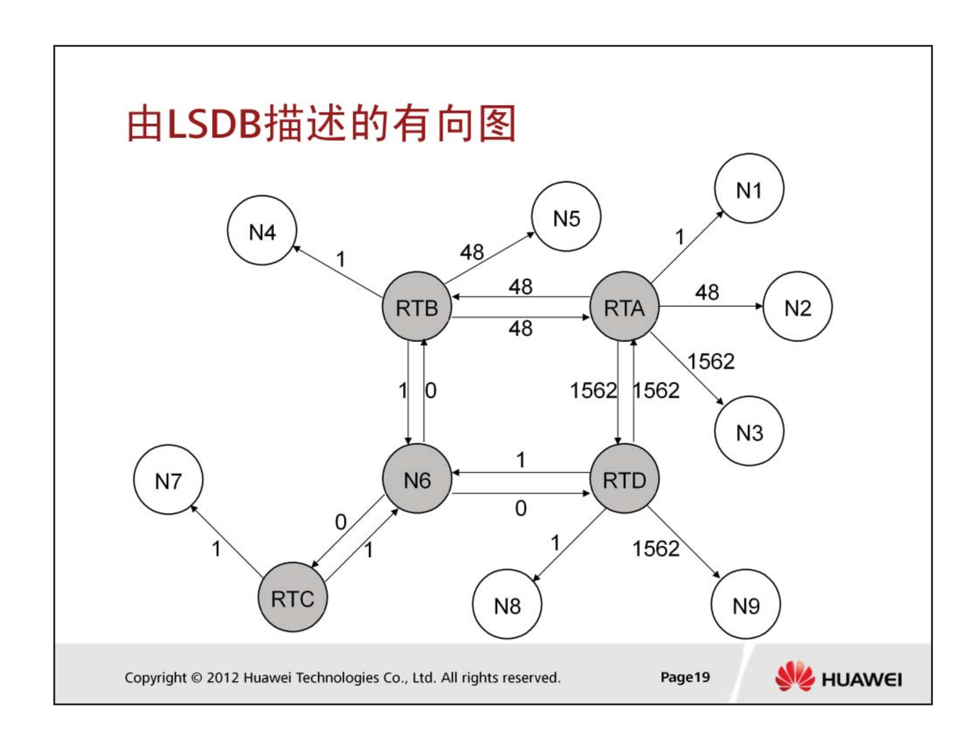
每一个路由器的Router ID为该路由器Loopback0的IP地址。

红色数字表示接口开销。



这是Area 1的LSDB。

在该LSDB中，有四条Router-LSA分别描述四台路由器的活动连接，有一条Network-LSA描述网络中的广播型网段。



由LSDB描述的有向图如图所示。

有向图中有五个Transit节点，Loopback接口被描述成Stub网段。

计算过程的两个阶段

第一阶段	计算Transit节点，忽略Stub节点，生成一个最短路径树
第二阶段	只计算Stub节点，将Stub网段挂到最短路径树上去

Page20

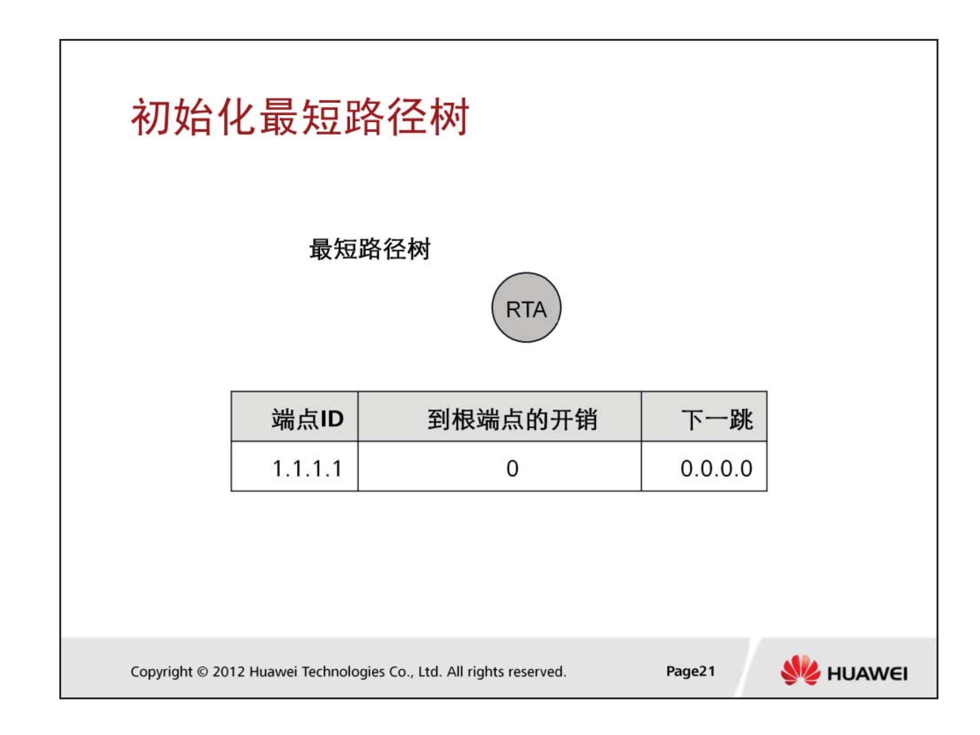


每个路由器计算以自己为根的最短路径树。

计算最短路径树的过程分为两个阶段：

第一阶段，计算所有的Transit节点，包括路由器和Transit网段。

第二阶段，计算Stub网段。



端点ID	到根端点的开销	下一跳
1.1.1.1	0	0.0.0.0

本例中，解释RTA如何计算以RTA为根的最短路径树。

计算过程中首先初始化最短路径树，RTA将自己做为根节点添加到最短路径树上。

计算最短路径树－计算描述RTA的LSA

```
Type        : Router
Ls id       : 1.1.1.1
Adv rtr     : 1.1.1.1
Ls age      : 36
Len         : 84
Options     :  E
seq#        : 8000000b
chksum      : 0xc250
Link count: 5
    Link ID: 2.2.2.2
    Data    : 10.1.1.1
    Link Type: P-2-P
    Metric : 48
    Link ID: 10.1.1.0
    Data    : 255.255.255.252
    Link Type: StubNet
    Metric : 48
    Link ID: 4.4.4.4
    Data    : 10.5.1.1
    Link Type: P-2-P
    Metric : 1562
    Link ID: 10.5.1.0
    Data    : 255.255.255.252
    Link Type: StubNet
    Metric : 1562
    Link ID: 1.1.1.1
    Data    : 255.255.255.255
    Link Type: StubNet
    Metric : 1
```

候选列表

端点ID	到根端点的开销	下一跳
4.4.4.4	1562	10.4.1.1
2.2.2.2	48	10.2.1.1

 Page22 HUAWEI

RTA将自己添加到最短路径树上之后，检查自己生成的Router-LSA，对于该LSA中所描述的每一个连接，如果不是一个Stub连接，就把该连接添加到候选列表中，端点ID为Link ID，到根端点的开销为LSA中描述的Metric值。本例中，添加端点4.4.4.4和2.2.2.2。

计算最短路径树—计算描述RTA的LSA

最短路径树

端点ID	到根端点的开销	下一跳
1.1.1.1	0	0.0.0.0
2.2.2.2	48	10.2.1.1

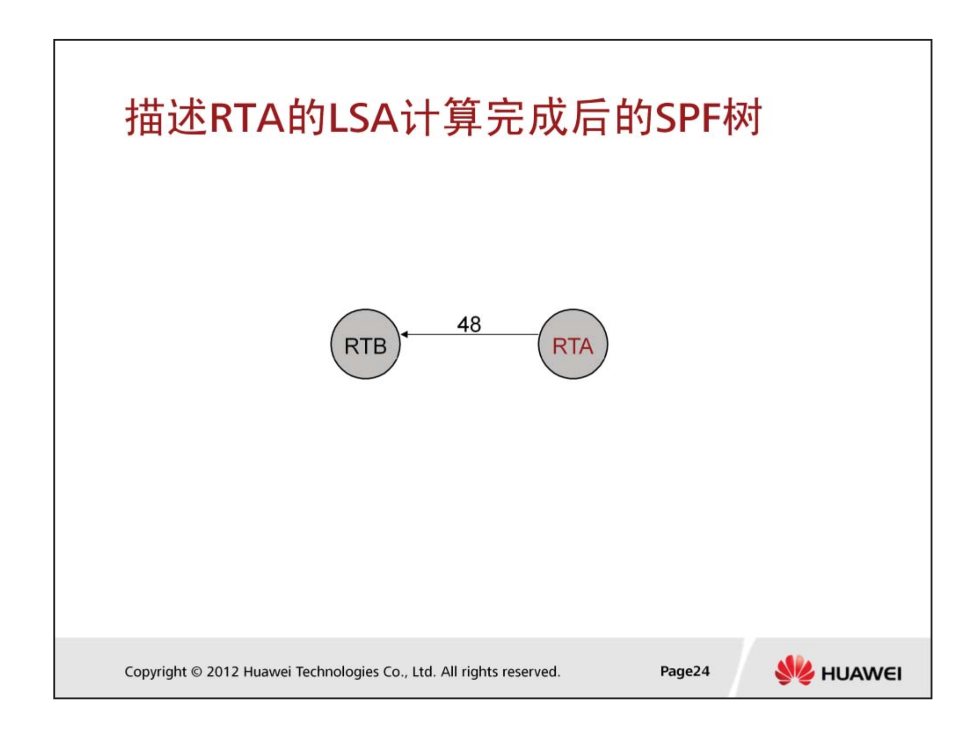
候选列表

端点ID	到根端点的开销	下一跳
4.4.4.4	1562	10.4.1.1

 Page23



将候选列表中到根端点开销最小的端点移到最短路径树上。本例中，将2.2.2.2移到最短路径树上。



上图为使用有向图描述此时的最短路径树，RTB为新添加的节点。

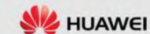
计算最短路径树－计算描述RTB的LSA

```
Type        : Router
Ls id       : 2.2.2.2
Adv rtr     : 2.2.2.2
Ls age      : 470
Len         : 72
Options     :  E
seq#        : 8000000e
chksum      : 0xbfa4
Link count: 4
   Link ID: 1.1.1.1
   Data    : 10.2.1.1
   Link Type: P-2-P
   Metric : 48
   Link ID: 10.2.1.0
   Data    : 255.255.255.252
   Link Type: StubNet
   Metric : 48
   Link ID: 10.3.1.1
   Data    : 10.3.1.1
   Link Type: TransNet
   Metric : 1
   Link ID: 2.2.2.2
   Data    : 255.255.255.255
   Link Type: StubNet
   Metric : 1
```

候选列表

端点ID	到根端点的开销	下一跳
4.4.4.4	1562	10.4.1.1
10.3.1.1	48+1=49	10.2.1.1

　Page25



当有新节点添加到最短路径树上的时候，则检查LS ID为新节点的端点ID的LSA，本例中检查LS ID为2.2.2.2的LSA。

如果LSA中所描述的连接的Link ID在最短路径树上已经存在，则忽略该连接。本例中，Link ID为1.1.1.1的连接被忽略，只有10.3.1.1的连接被添加到候选列表中。到根端点的开销设置为此连接的Metric值（本例中此连接的Metric值为1）与父端点（本例中此连接的父端点为2.2.2.2）到根端点的开销（本例中此开销值为48）之和。

计算最短路径树—计算描述RTB的LSA

最短路径树

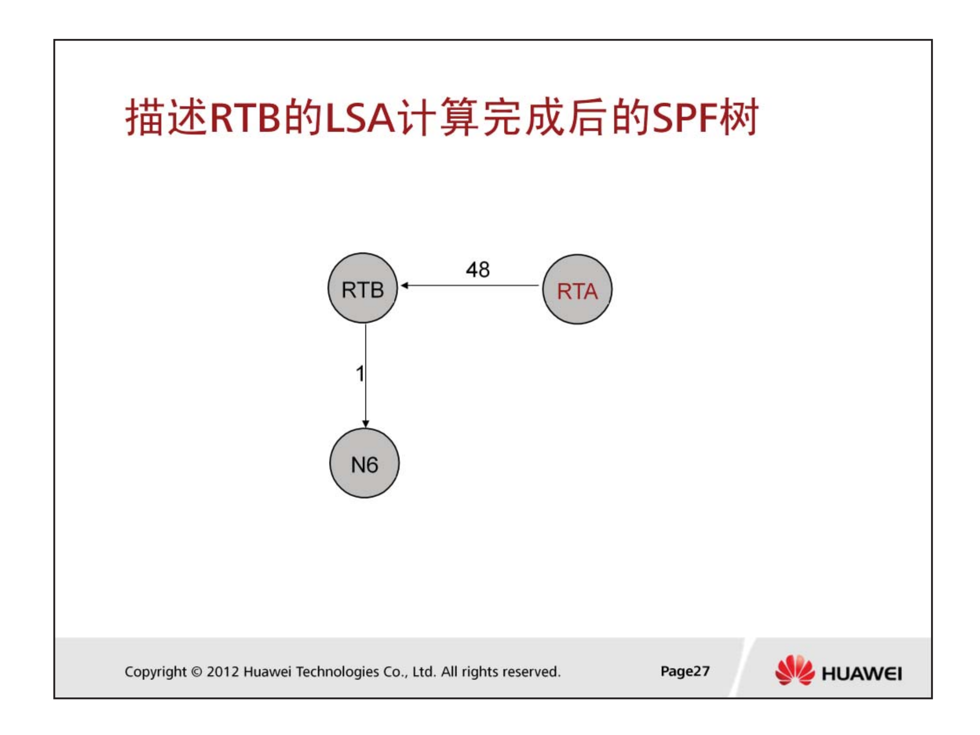
端点ID	到根端点的开销	下一跳
1.1.1.1	0	0.0.0.0
2.2.2.2	48	10.2.1.1
10.3.1.1	49	10.2.1.1

候选列表

端点ID	到根端点的开销	下一跳
4.4.4.4	1562	10.4.1.1

 Page26 

将候选列表中到根端点的开销最小的端点移动到最短路径树上，本例中，将10.3.1.1移到最短路径树上。



上图为使用有向图描述此时的最短路径树，N6为新添加到最短路径树上的节点。

计算最短路径树—计算描述10.3.1.0/24的LSA

```
Type        : Network
Ls id       : 10.3.1.1
Adv rtr     : 2.2.2.2
Ls age      : 811
Len         : 36
Options     :  E
seq#        : 80000007
chksum      : 0x39db
Net mask    : 255.255.255.0
  Attached Router    3.3.3.3
  Attached Router    4.4.4.4
  Attached Router    2.2.2.2
```

候选列表

端点ID	到根端点的开销	下一跳
4.4.4.4	1562	10.4.1.1
3.3.3.3	49+0=49	10.2.1.1
4.4.4.4	49+0=49	10.2.1.1

 Page28



检查LS ID为最新添加节点的端点ID的LSA，本例中检查LS ID为10.3.1.1的LSA。

在所描述的连接中，忽略2.2.2.2，将3.3.3.3和4.4.4.4添加到候选列表中。从Transit网段到所连路由器的开销为0。

如果在候选列表中出现两个端点ID一样但是到根端点的开销不一样的端点，则删除到根端点的开销大的端点。

计算最短路径树—计算描述 10.3.1.0/24的LSA

最短路径树

端点ID	到根端点的开销	下一跳
1.1.1.1	0	0.0.0.0
2.2.2.2	48	10.2.1.1
10.3.1.1	49	10.2.1.1
3.3.3.3	49	10.2.1.1

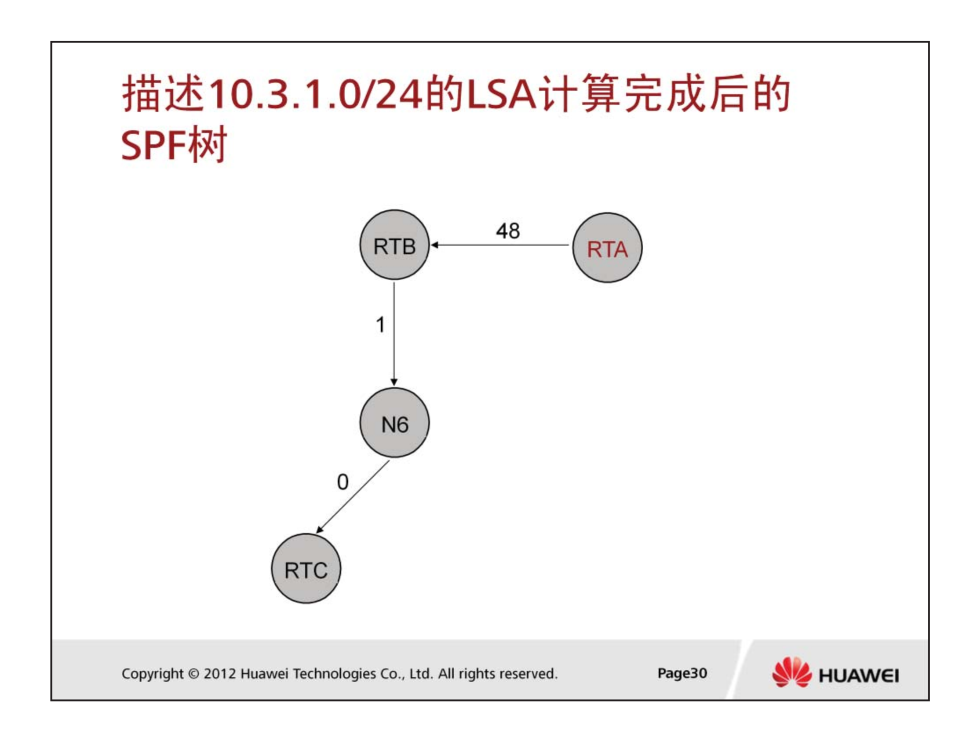
候选列表

端点ID	到根端点的开销	下一跳
4.4.4.4	49+0=49	10.2.1.1

 Page29



将候选列表中到根端点的开销最小的端点移动到最短路径树上，本例中，将3.3.3.3移到最短路径树上。



上图为使用有向图描述此时的最短路径树，RTC为新添加到最短路径树上的节点。

计算最短路径树—计算描述RTC的LSA

```
Type       : Router
Ls id      : 3.3.3.3
Adv rtr    : 3.3.3.3
Ls age     : 1015
Len        : 48
Options    :  E
seq#       : 8000000a
chksum     : 0x886d
Link count: 2
   Link ID: 10.3.1.1
   Data    : 10.3.1.2
   Link Type: TransNet
   Metric : 1
   Link ID: 3.3.3.3
   Data    : 255.255.255.255
   Link Type: StubNet
   Metric : 1
```

候选列表

端点ID	到根端点的开销	下一跳
4.4.4.4	49+0=49	10.2.1.1

 Page31



检查LS ID为最新添加节点的端点ID的LSA，本例中检查LS ID为3.3.3.3的LSA。

本例中，没有新端点被添加到候选列表中。

计算最短路径树－计算描述RTC的LSA

最短路径树

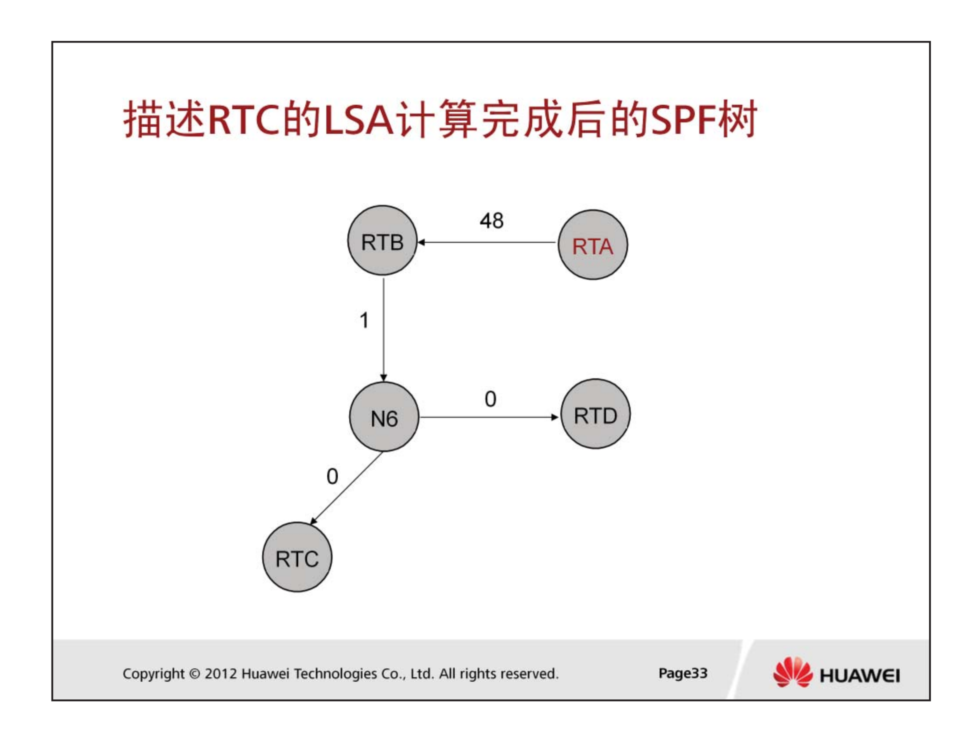
端点ID	到根端点的开销	下一跳
1.1.1.1	0	0.0.0.0
2.2.2.2	48	10.2.1.1
10.3.1.1	49	10.2.1.1
3.3.3.3	49	10.2.1.1
4.4.4.4	49	10.2.1.1

候选列表

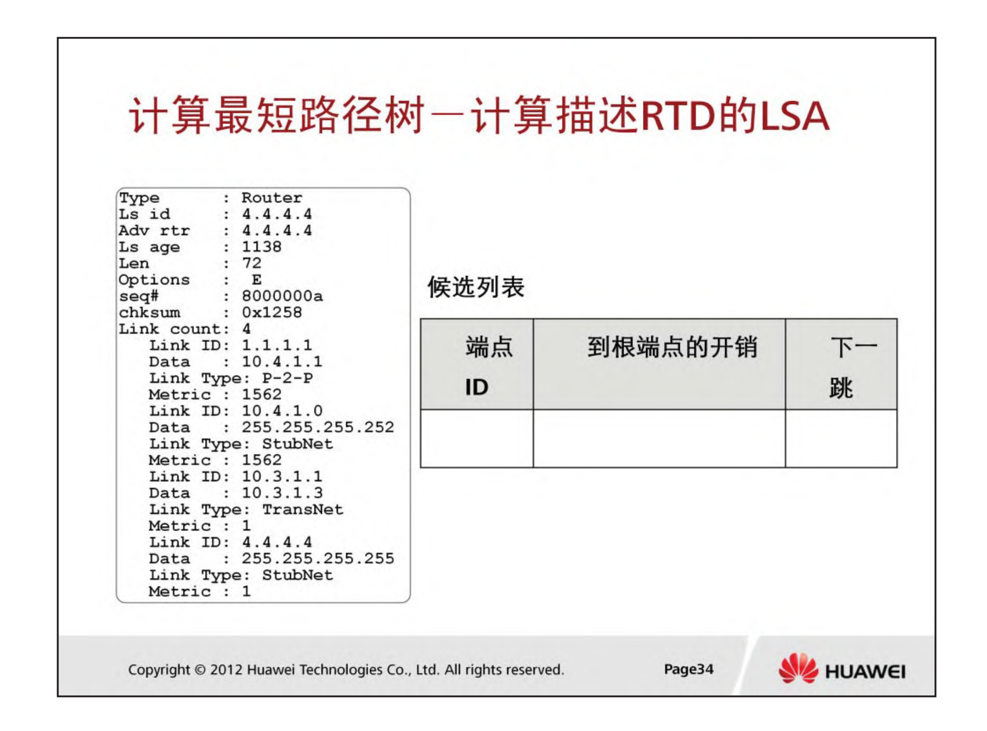
端点ID	到根端点的开销	下一跳

 Page32 HUAWEI

将候选列表中到根端点的开销最小的端点移动到最短路径树上，本例中，将4.4.4.4移到最短路径树上。



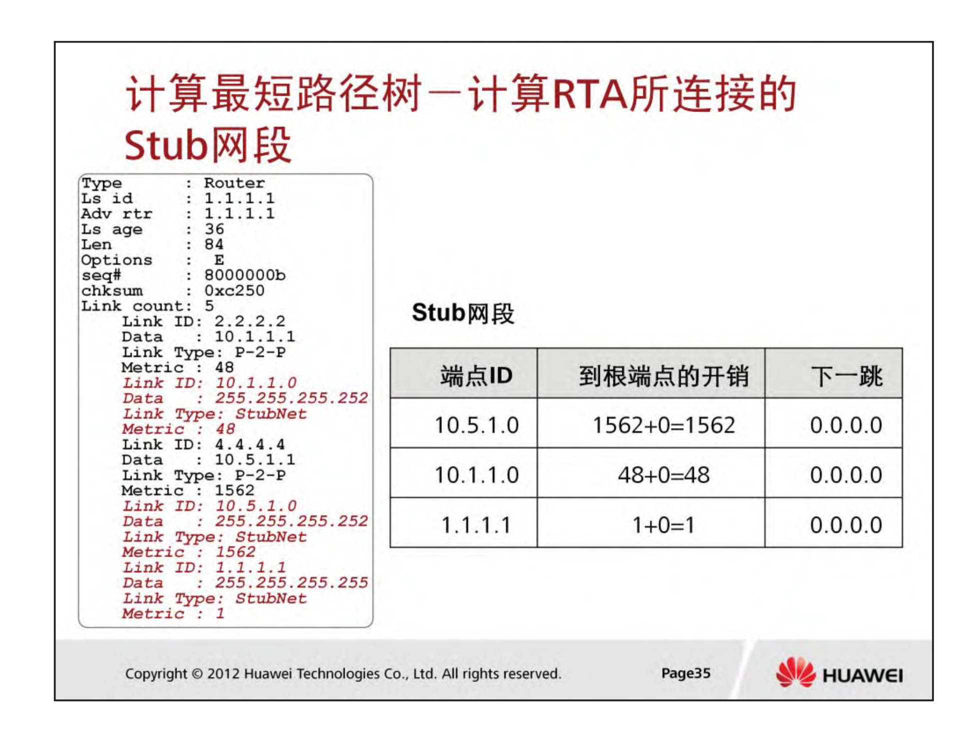
上图为使用有向图描述此时的最短路径树，RTD为新添加到最短路径树上的节点。



检查LS ID为最新添加节点的端点ID的LSA，本例中检查LS ID为4.4.4.4的LSA。

本例中，没有新端点被添加到候选列表中。

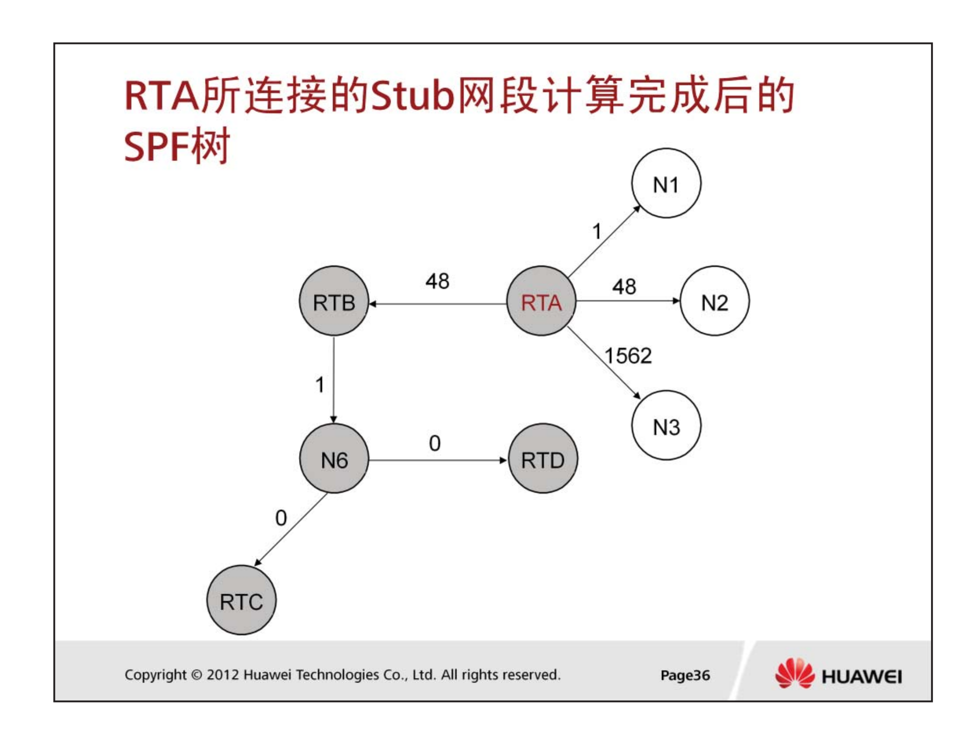
如果在此时候选列表为空，则计算最短路径树的第一阶段结束。



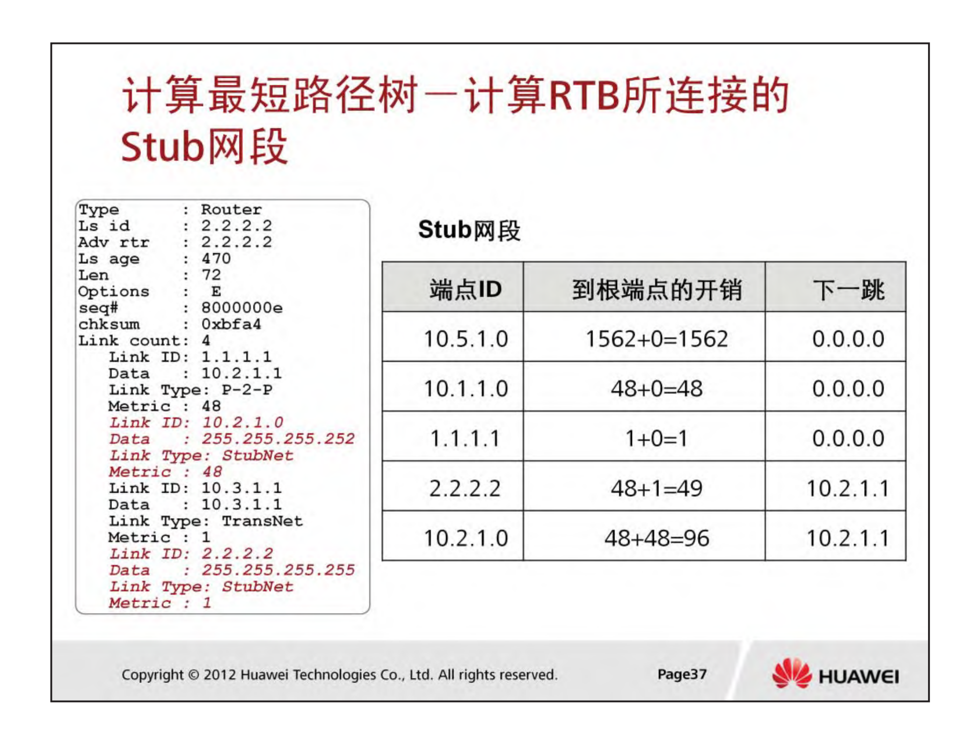
端点ID	到根端点的开销	下一跳
10.5.1.0	1562+0=1562	0.0.0.0
10.1.1.0	48+0=48	0.0.0.0
1.1.1.1	1+0=1	0.0.0.0

检查每个路由器端点的Router-LSA，计算Stub网段。

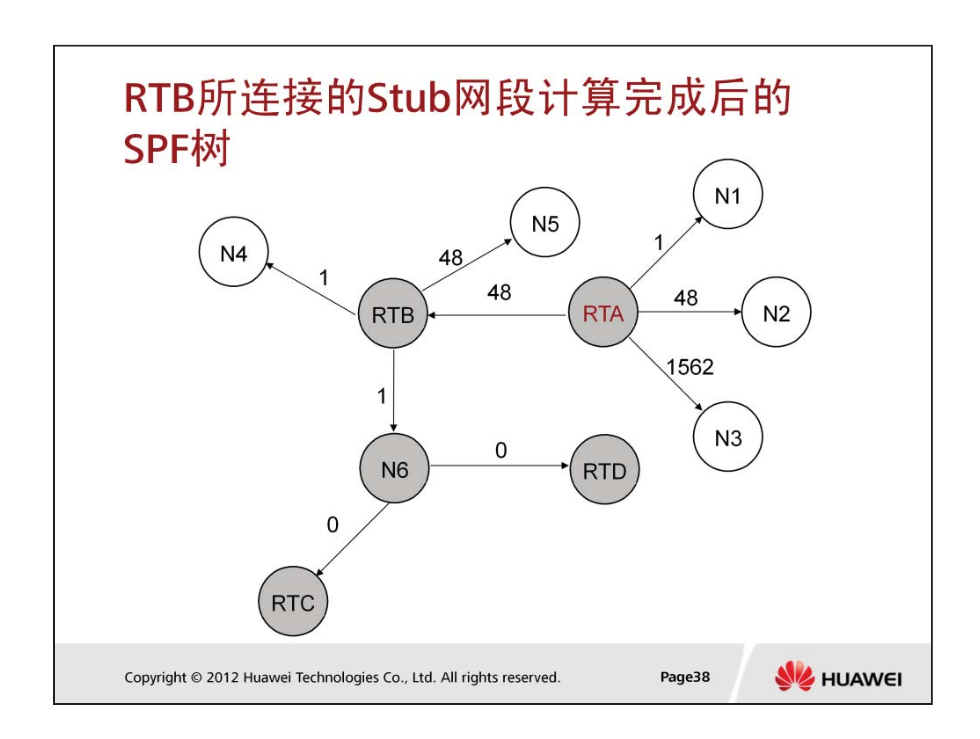
本例中，首先检查RTA的Router-LSA，共有三个Stub网段。



上图为使用有向图描述此时的最短路径树，三个Stub网段N1、N2、N3为新添加到最短路径树上的节点。



检查RTB的Router-LSA。共有两个Stub网段。



上图为使用有向图描述此时的最短路径树，两个Stub网段N4、N5为新添加到最短路径树上的节点。

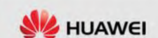
计算最短路径树－计算RTC所连接的Stub网段

```
Type       : Router
Ls id      : 3.3.3.3
Adv rtr    : 3.3.3.3
Ls age     : 1015
Len        : 48
Options    :  E
seq#       : 8000000a
chksum     : 0x886d
Link count: 2
   Link ID: 10.3.1.1
   Data   : 10.3.1.2
   Link Type: TransNet
   Metric : 1
   Link ID: 3.3.3.3
   Data   : 255.255.255.255
   Link Type: StubNet
   Metric : 1
```

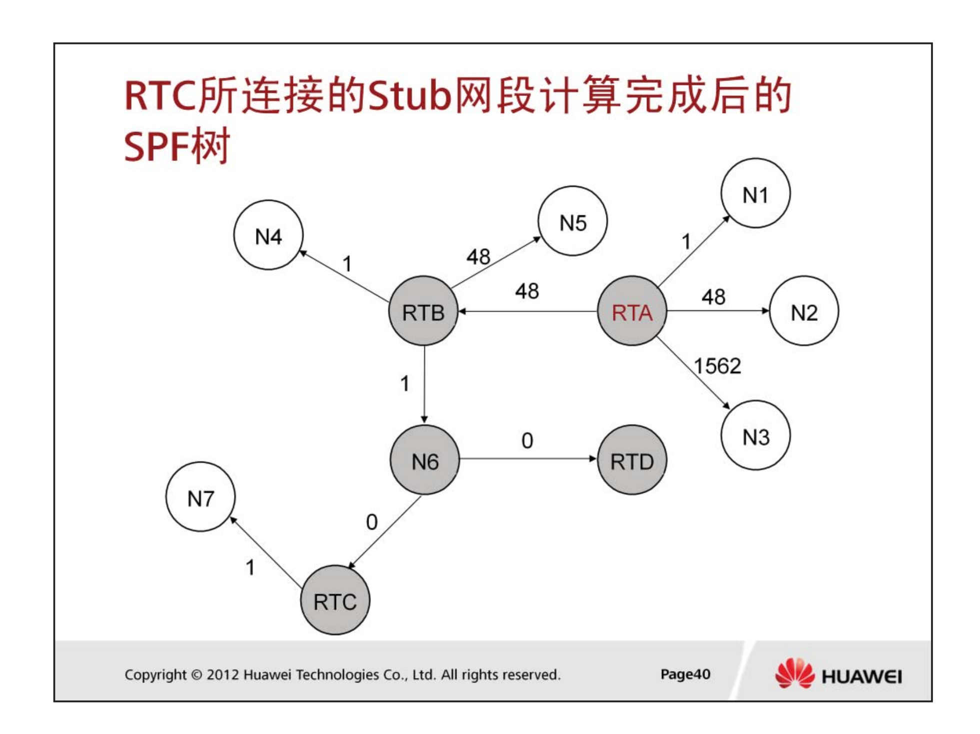
Stub网段

端点ID	到根端点的开销	下一跳
10.5.1.0	1562+0=1562	0.0.0.0
10.1.1.0	48+0=48	0.0.0.0
1.1.1.1	1+0=1	0.0.0.0
2.2.2.2	48+1=49	10.2.1.1
10.2.1.0	48+48=96	10.2.1.1
3.3.3.3	49+1=50	10.2.1.1

 Page39



检查RTC的Router-LSA。只有一个Stub网段。



上图为使用有向图描述此时的最短路径树，Stub网段N7为新添加到最短路径树上的节点。

计算最短路径树—计算RTD所连接的Stub网段

```
Type        : Router
Ls id       : 4.4.4.4
Adv rtr     : 4.4.4.4
Ls age      : 1138
Len         : 72
Options     :  E
seq#        : 8000000a
chksum      : 0x1258
Link count: 4
   Link ID: 1.1.1.1
   Data    : 10.4.1.1
   Link Type: P-2-P
   Metric : 1562
   Link ID: 10.4.1.0
   Data    : 255.255.255.252
   Link Type: StubNet
   Metric : 1562
   Link ID: 10.3.1.1
   Data    : 10.3.1.3
   Link Type: TransNet
   Metric : 1
   Link ID: 4.4.4.4
   Data    : 255.255.255.255
   Link Type: StubNet
   Metric : 1
```

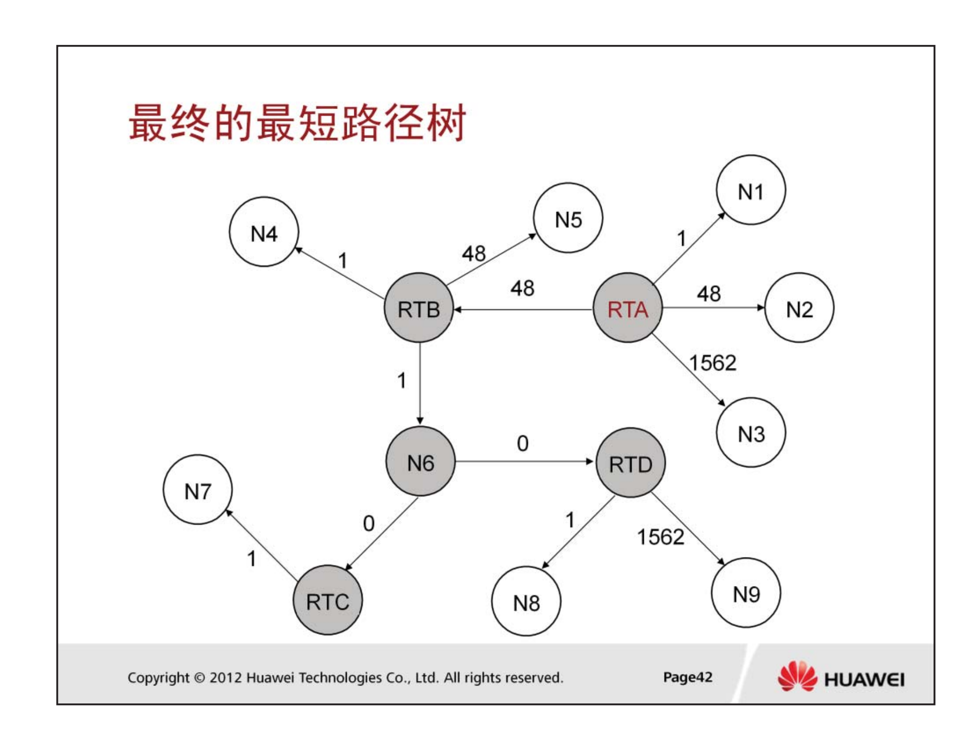
Stub网段

端点ID	到根接口的开销	下一跳
10.5.1.0	1562+0=1562	0.0.0.0
10.1.1.0	48+0=48	0.0.0.0
1.1.1.1	1+0=1	0.0.0.0
2.2.2.2	48+1=49	10.2.1.1
10.2.1.0	48+48=96	10.2.1.1
3.3.3.3	49+1=50	10.2.1.1
4.4.4.4	49+1=50	10.2.1.1
10.4.1.0	49+1562=1611	10.2.1.1

 Page41 HUAWEI

检查RTD的Router-LSA。共有两个Stub网段。

最短路径树和Stub网段的列表给出了到达网络中所有目的地的路由。



上图为使用有向图描述最终的最短路径树，两个Stub网段N8、N9为新添加到最短路径树上的节点。



问 题

Router-LSA描述的链接类型有哪些?

Network-LSA中除了LSA头部还有哪些信息?

计算最短路径树分几个阶段?

　Page43



Router-LSA描述的链接类型有哪些?

Router-LSA描述的链接类型有四种，分别是点到点链接，Transit链接，Stub链接和虚连接。

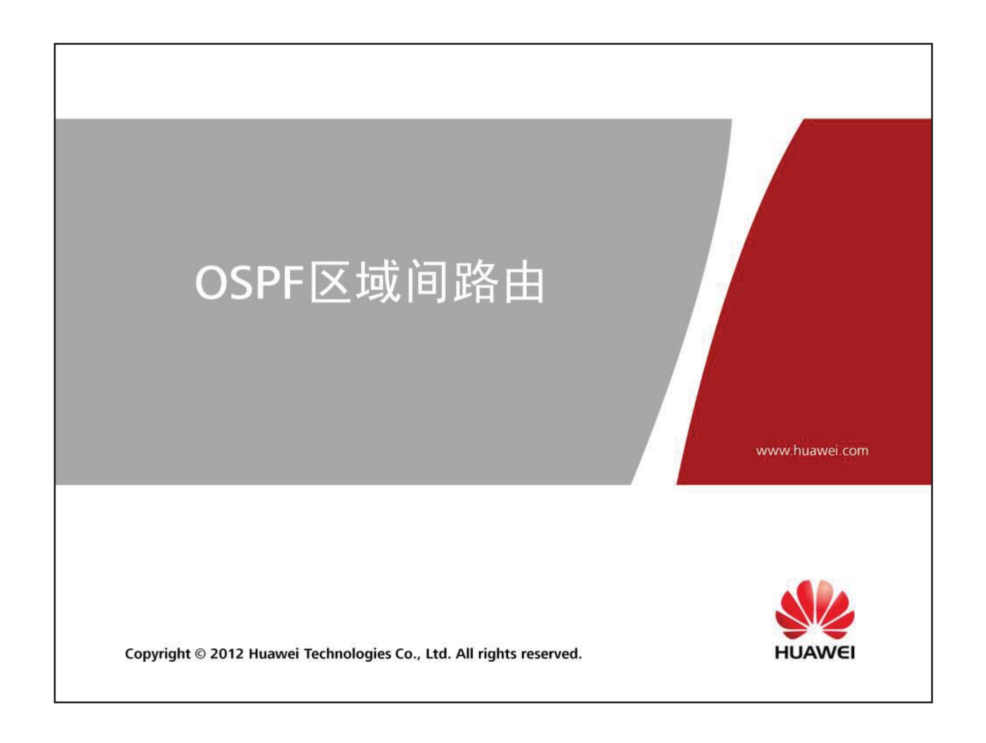
Network-LSA中除了LSA头部还有哪些重要信息?

所描述网段的网络掩码和该网段上所连接路由器的列表。

计算最短路径树分几个阶段?

两个阶段，第一阶段计算路由器节点和Transit网段；第二阶段计算Stub网段。


谢谢
www.huawei.com


OSPF区域间路由
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍OSPF区域间路由技术。

课程内容包括区域间路由原理，使用Network-Summary-LSA描述区域间路由信息，虚连接技术，区域间路由汇聚等内容。

Page1



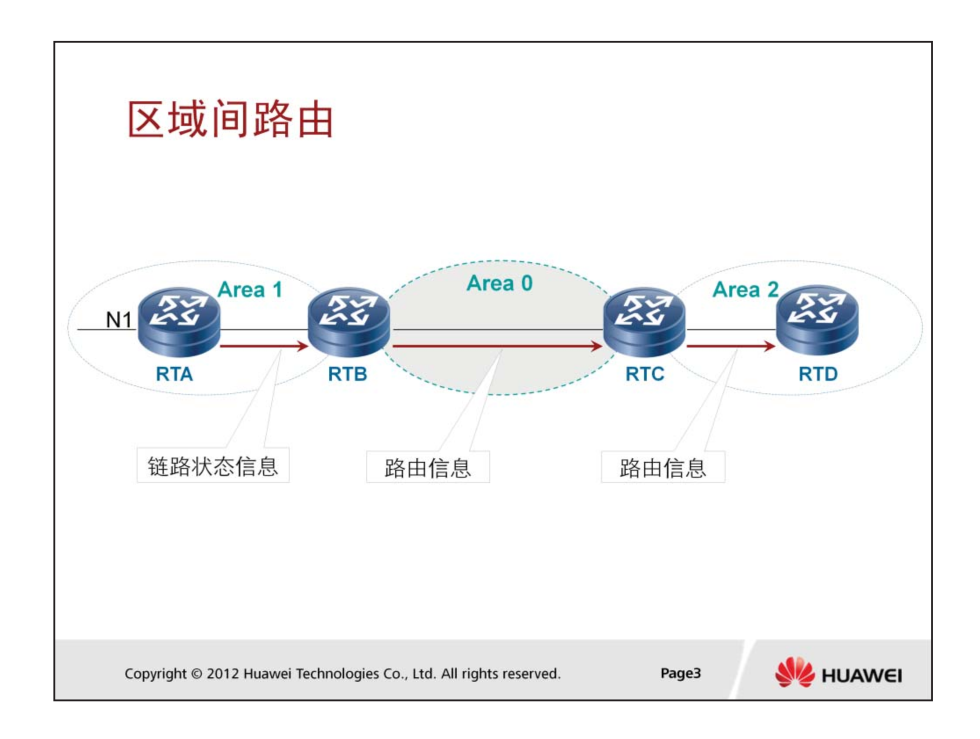


学完本课程后，您应该能：

- 理解区域间路由原理
- 理解Network-Summary-LSA
- 理解虚连接
- 理解区域间路由汇聚

Page2





区域边界路由器（ABR）上有多个LSDB，ABR为每一个区域维护一个LSDB。

ABR将所连接的非骨干区域内的链路状态信息抽象成路由信息，并发布到骨干区域中，由骨干区域进一步发布到其他非骨干区域中。

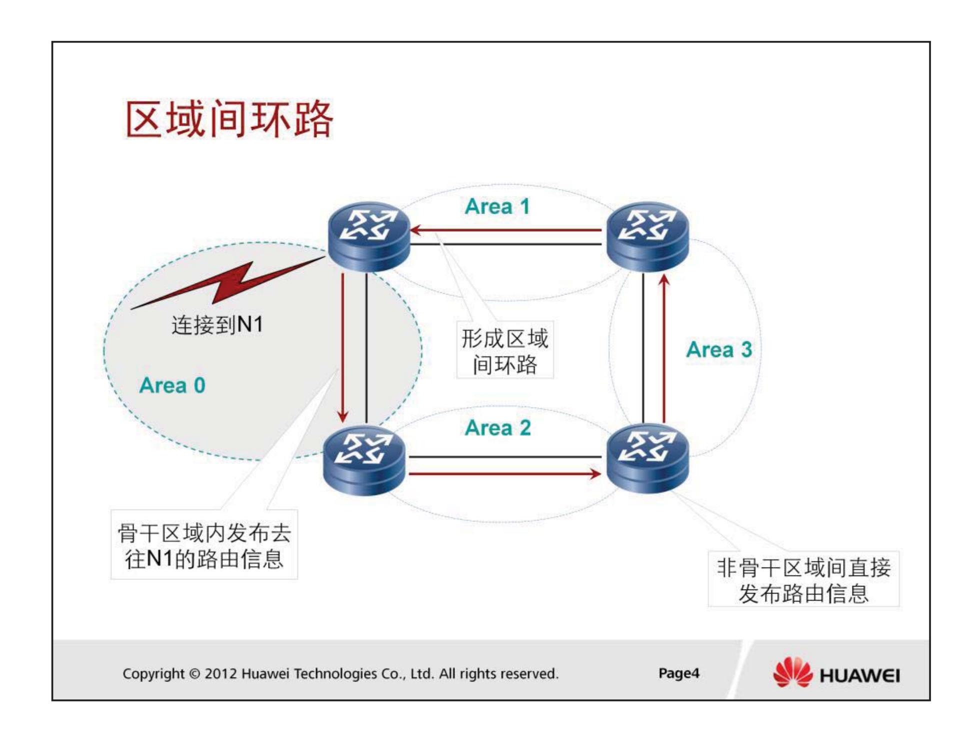
ABR也要将骨干区域的链路状态信息抽象成路由信息，并发布到所连接的非骨干区域中。

本例中:

RTA生成关于N1的链路状态信息并泛洪到RTB。

RTB生成关于N1的抽象路由信息并在骨干区域内泛洪。

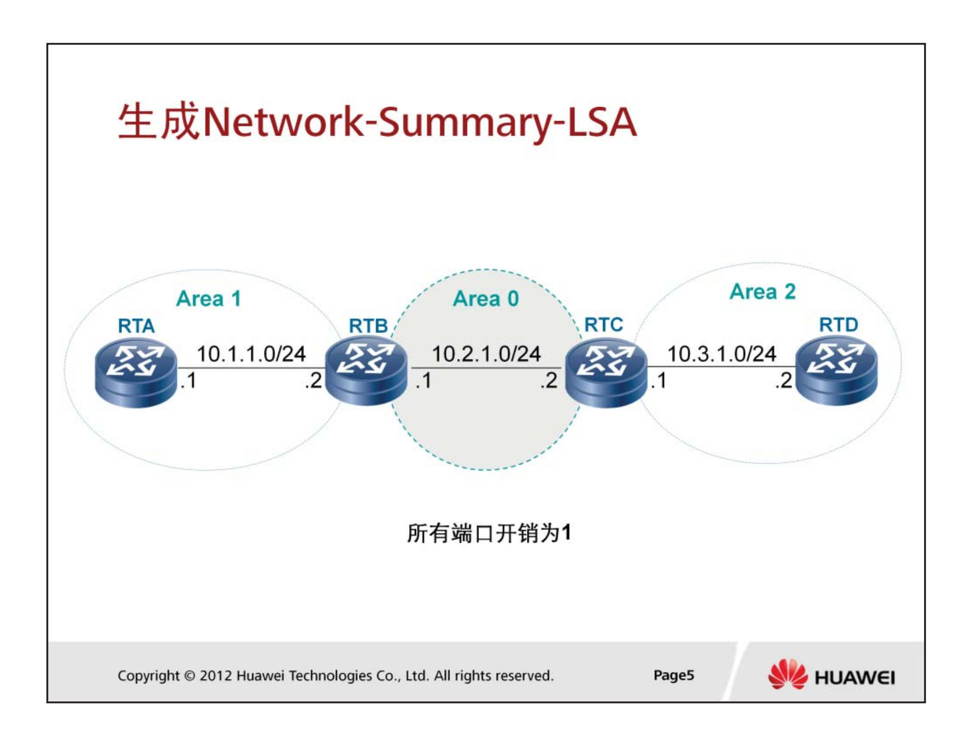
RTC将接收到的抽象路由信息泛洪到RTD。



本例中:

如果直接在Area 2和Area 3之间发布路由信息是允许的，那么一个区域间的环路就会形成。

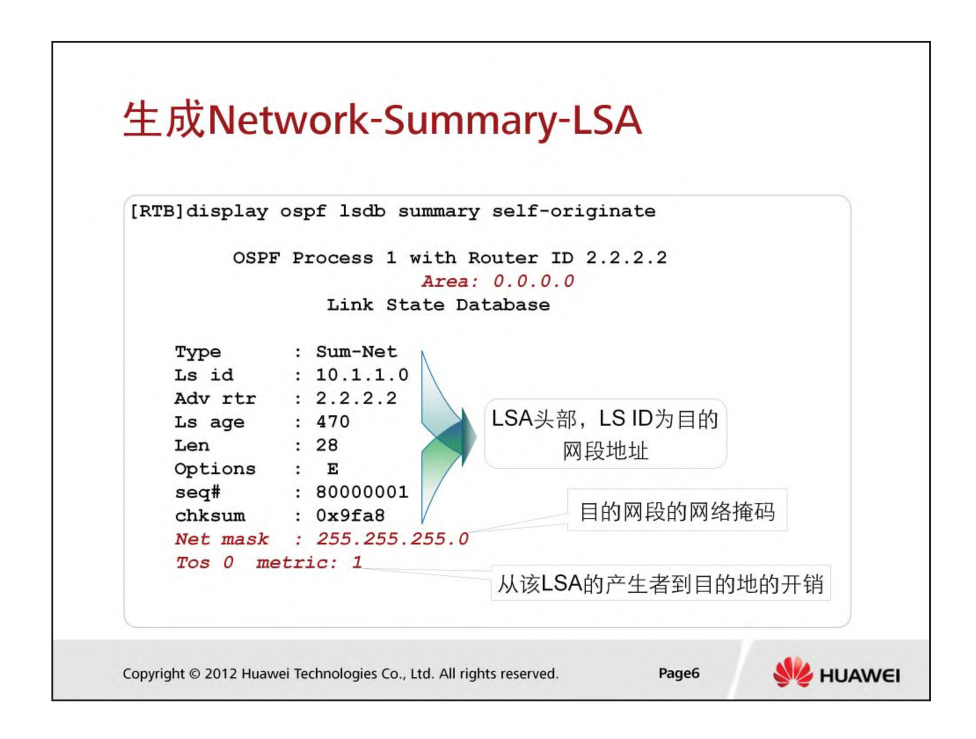
为了避免区域间的环路，OSPF规定不允许直接在两个非骨干区域之间发布路由信息，只允许在一个区域内部或者在骨干区域和非骨干区域之间发布路由信息。因此，每个区域边界路由器（ABR）都必须连接到骨干区域。



本例中:

RTA的Router ID为1.1.1.1，RTB为2.2.2.2，RTC为3.3.3.3，RTD为4.4.4.4
。

所有端口的开销为1。



Network-Summary-LSA中主要包括以下内容：

Link State ID被设置成目的网段的IP地址。

Net mask被设置成目的网段的网络掩码。

Metric被设置成从该ABR到达目的网段的开销值。

以网段10.1.1.0/24为例，区域间路由发布的过程如下：

首先，RTB（Area 1的ABR）将该网段的路由信息发布到骨干区域中。

生成Network-Summary-LSA

```
[RTC]display ospf lsdb summary

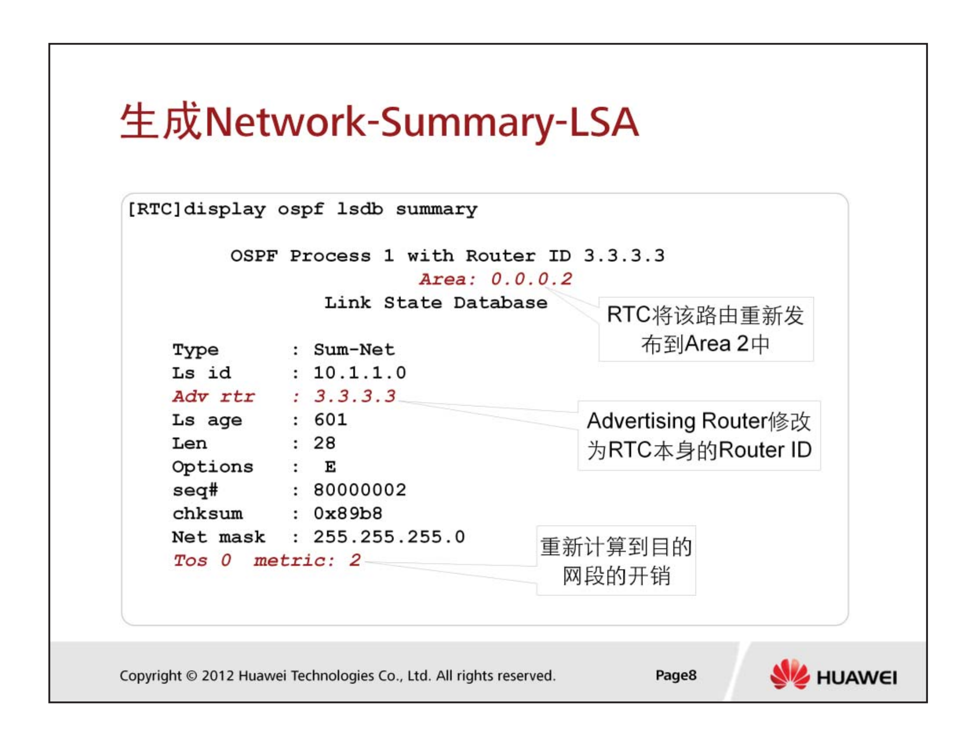
        OSPF Process 1 with Router ID 3.3.3.3
                        Area: 0.0.0.0
                Link State Database

    Type        : Sum-Net
    Ls id       : 10.1.1.0
    Adv rtr     : 2.2.2.2
    Ls age      : 705
    Len         : 28
    Options     :  E
    seq#        : 80000002
    chksum      : 0x9da9
    Net mask    : 255.255.255.0
    Tos 0  metric: 1
```

RTC通过骨干区域学习到RTB发布的路由

 Page7 HUAWEI

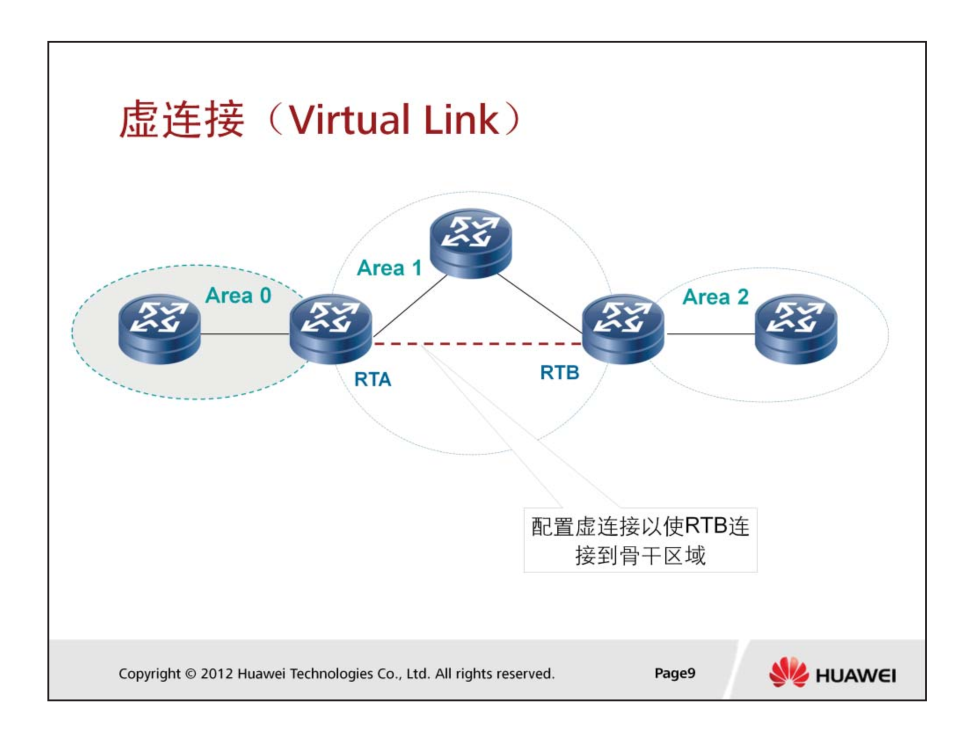
然后，RTC通过骨干区域学习到RTB发布的关于网段10.1.1.0/24的路由信息。



最后，RTC根据从骨干区域学习到Network-Summary-LSA重新生成一条新的Network-Summary-LSA，并发布到Area 2中，在这条新的LSA中：

Advertising Router修改为RTC本身的Router ID；

到目的网段的开销需要重新计算，修改为从RTC到目的网段的总开销。

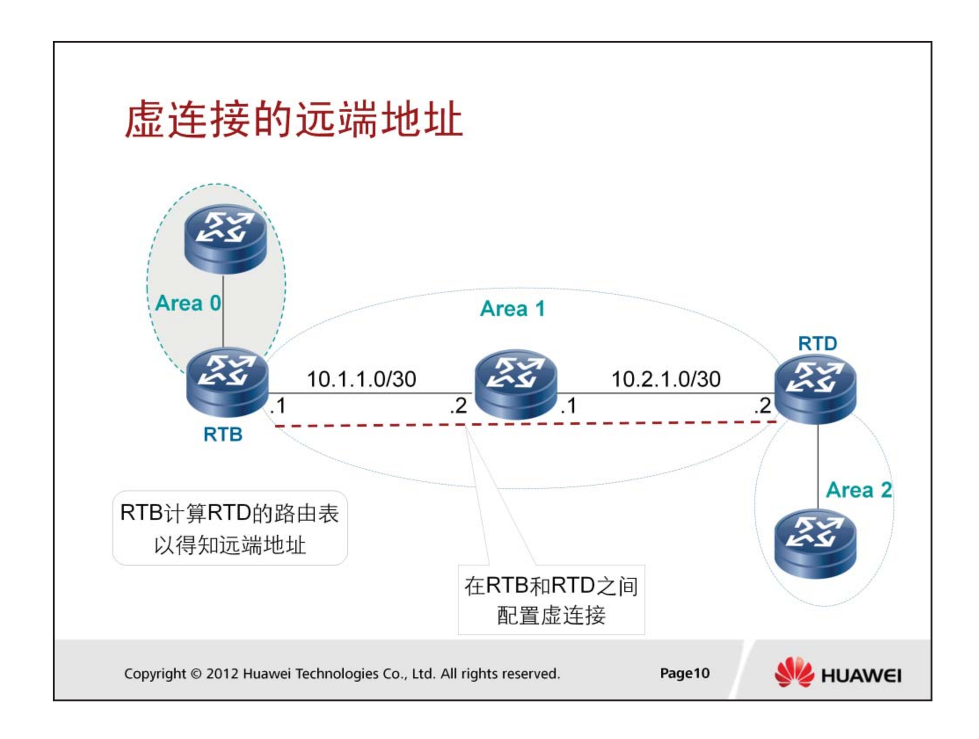


骨干区域必须是连续的，但是并不要求物理上连续，可以使用虚连接使骨干区域逻辑上连续。

虚连接可以在任意两个区域边界路由器上建立，但是要求这两个区域边界路由器都有端口连接到一个共同的非骨干区域。

虚连接是属于骨干区域（Area 0）的一条虚拟链路。

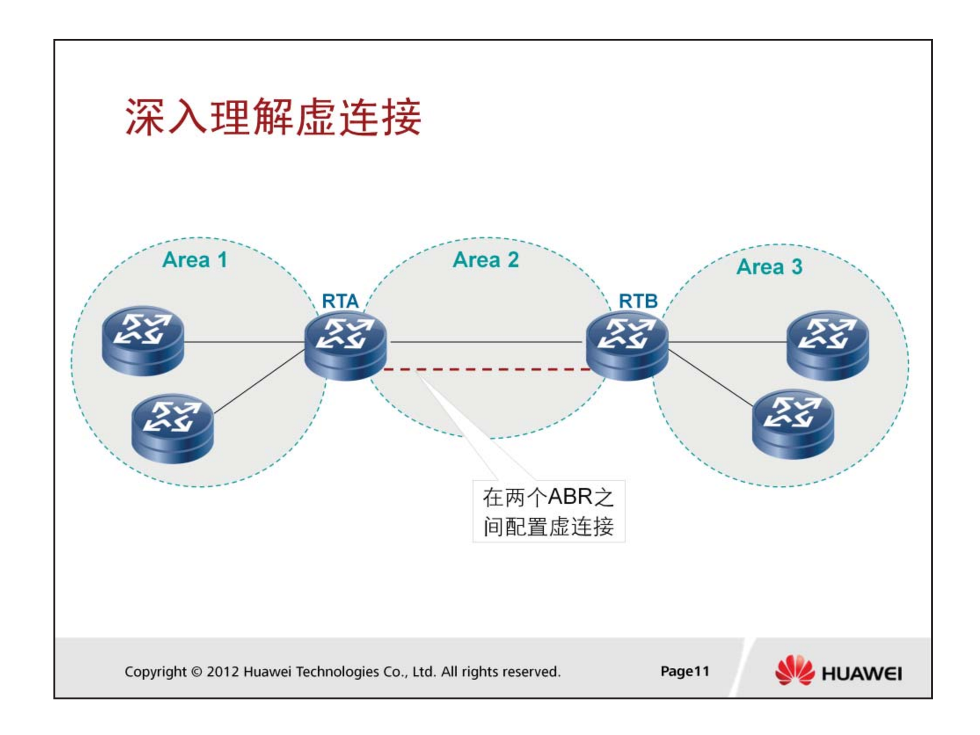
本例中，在RTA和RTB之间建立了一条虚连接，以使RTB连接到骨干区域。



虚连接的两个端点需要相互交换协议报文，但是虚连接的邻居是用邻居的Router ID来标识的，不能做为协议报文的目的IP地址。如何确定协议报文的目的IP地址呢?

每个虚连接的端点都要计算两个最短路径树，一个是本地最短路径树，而另一个是虚连接邻居的最短路径树。如图中虚连接端点之一RTB要以自己为根计算Area1内的最短路径树（即本地路径树）；另外RTB还要计算以其虚连接邻居RTD为根的Area1内的最短路径树（即虚连接邻居的最短路径树）。RTD类同。

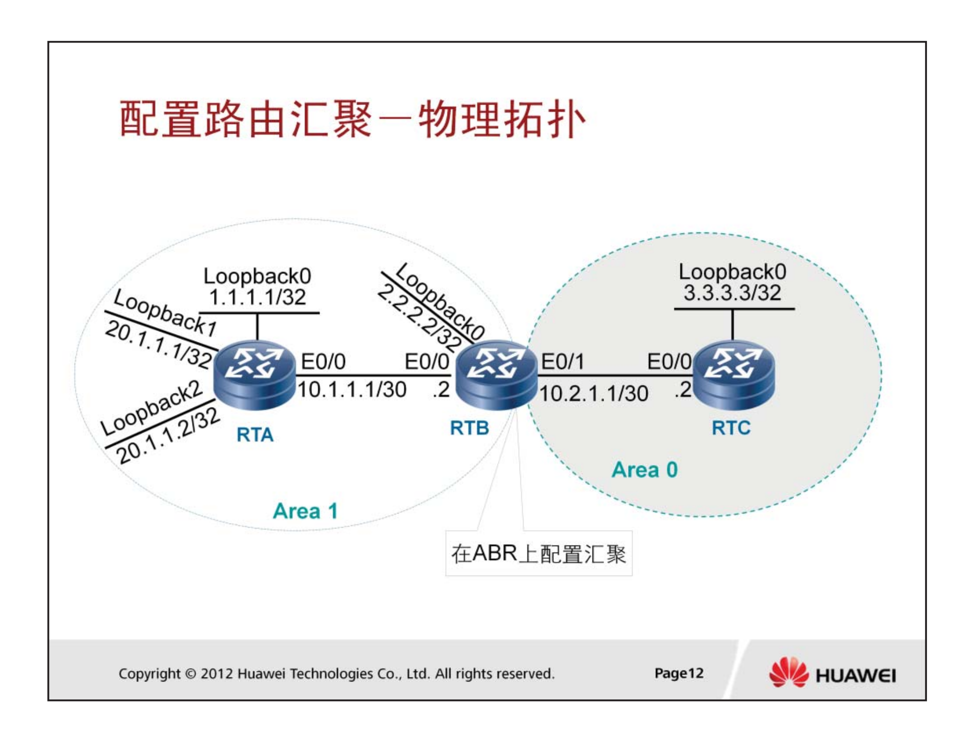
计算虚连接邻居的最短路径树之后，在最短路径树上查找如何到达本地路由器（通过Router ID标识），虚连接邻居到达本地路由器的出端口的IP地址为本地路由器发送给虚连接邻居的协议报文的目的IP地址。



本例中，网络中没有骨干区域（Area 0），怎样在这三个区域间发布路由信息?

只需要在两个区域边界路由器（RTA和RTB）之间配置一个虚连接即可。

虚连接是骨干区域的一部分，所有的虚连接都属于Area 0。



本例中:

RTB为ABR。

在RTB上，通过配置路由汇聚，把路由条目20.1.1.1/32和20.1.1.2/32汇聚成20.1.1.0/24。

汇聚之后，RTB通过Network-Summary-LSA向骨干区域发布路由信息的时候，只会发布汇聚之后的路由条目20.1.1.0/24，不发布明细路由条目。

配置路由汇聚—配置RTA和RTC

```
[RTA]router id 1.1.1.1
[RTA]ospf
[RTA-ospf-1]area 1
[RTA-ospf-1-area-0.0.0.1]network 1.1.1.1 0.0.0.0
[RTA-ospf-1-area-0.0.0.1]network 10.1.1.0 0.0.0.3
[RTA-ospf-1-area-0.0.0.1]network 20.1.1.1 0.0.0.0
[RTA-ospf-1-area-0.0.0.1]network 20.1.1.2 0.0.0.0
[RTA-ospf-1-area-0.0.0.1]return
<RTA>
```

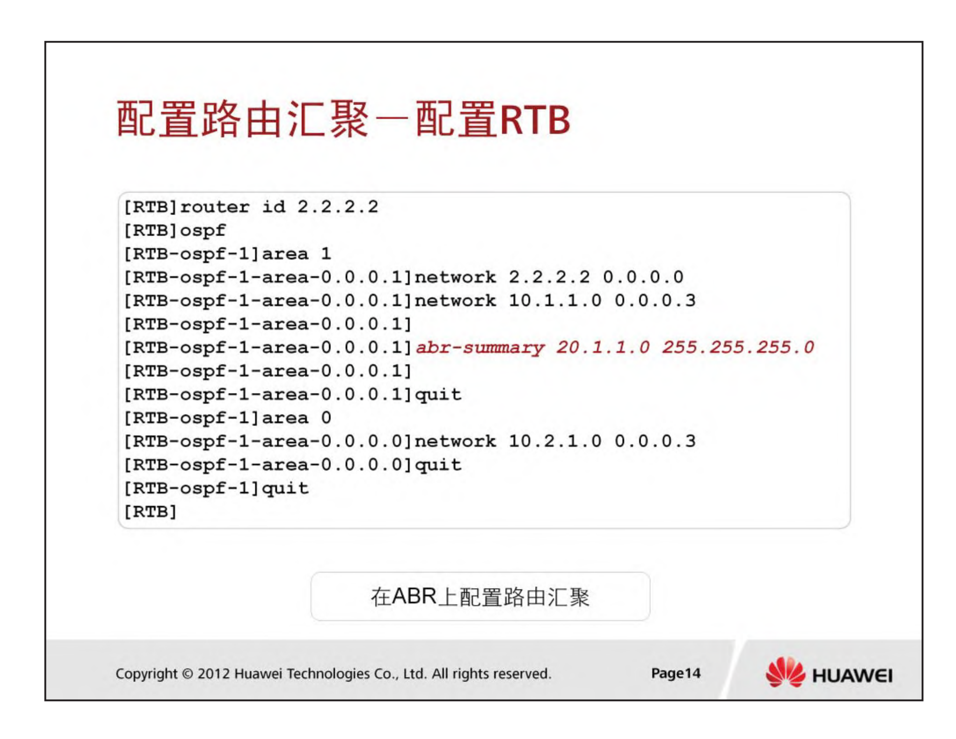
```
[RTC]router id 3.3.3.3
[RTC]ospf
[RTC-ospf-1]area 0
[RTC-ospf-1-area-0.0.0.0]network 3.3.3.3 0.0.0.0
[RTC-ospf-1-area-0.0.0.0]network 10.2.1.0 0.0.0.3
[RTC-ospf-1-area-0.0.0.0]return
<RTC>
```

 Page13 HUAWEI

把RTA的四个网段都配置在Area 1中。

在RTA上不配置路由汇聚，因为在RTA上配置路由汇聚是无意义的，RTA和RTB之间发布的是详细链路状态信息，而不是路由信息，而且RTA也不会生成Network-Summary-LSA。

配置RTC的两个网段属于Area 0。



在RTB的区域视图内（明细路由产生的区域）配置路由汇聚。

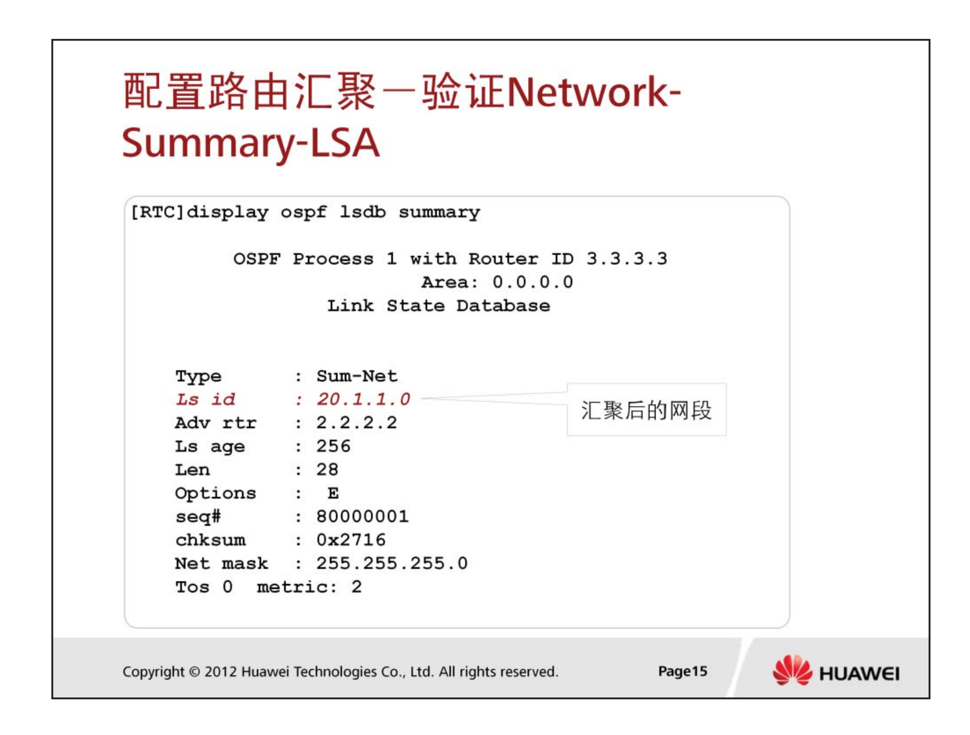
abr-summary ip-address mask [advertise | not-advertise] [cost cost]

advertise：将到这一聚合网段路由的摘要信息广播出去。

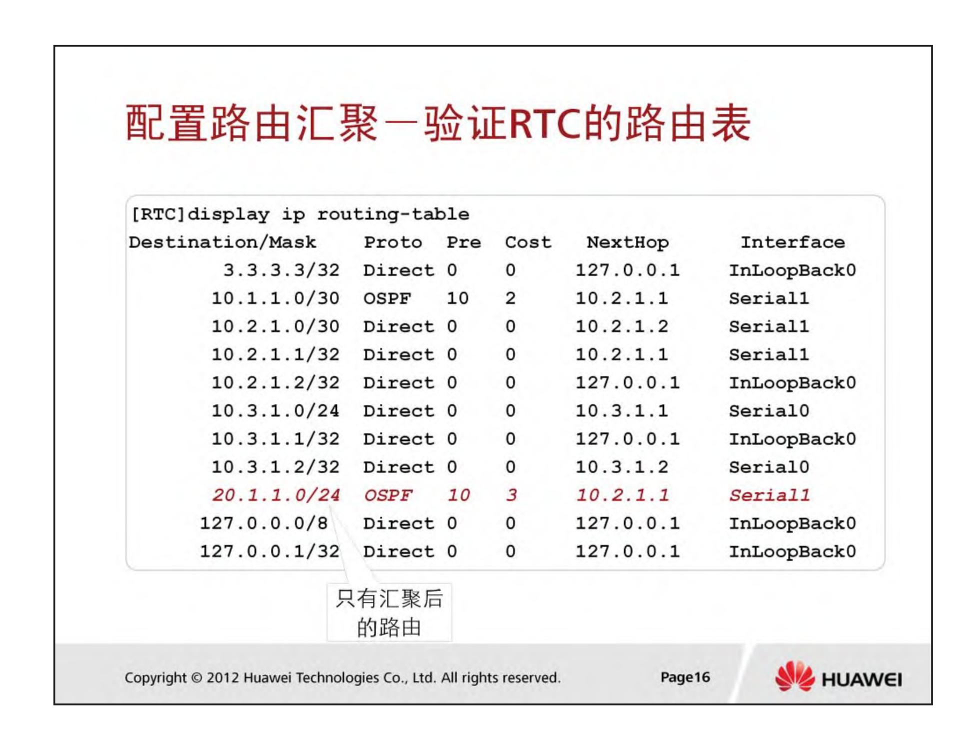
notadvertise：不将到这一聚合网段路由的摘要信息广播出去。

Cost：设置聚合路由的开销。

缺省情况下，只通告聚合之后的路由。



RTB在向骨干区域通告Area 1中的路由信息时，通过一条Network-Summary-LSA描述到汇聚之后的网段20.1.1.0/24的路由，不生成描述到网段20.1.1.1/32和20.1.1.2/32的明细路由的Network-Summary-LSA。



在RTC路由表中，只有聚合之后的路由20.1.1.0/24，没有明细路由20.1.1.1/32和20.1.1.2/32。



问 题

区域间传递的是否为链路状态信息?

如何避免区域间环路的形成?

如何确定虚连接的对端IP地址?

区域间路由汇聚功能在什么路由器上配置?

Page17



区域间传递的是否为链路状态信息?

不是，区域间传递的是路由信息，不是详细的链路状态信息。

如何避免区域间环路的形成?

只允许在骨干区域和非骨干区域之间发布路由信息，不允许在非骨干区域之间直接发布路由信息。

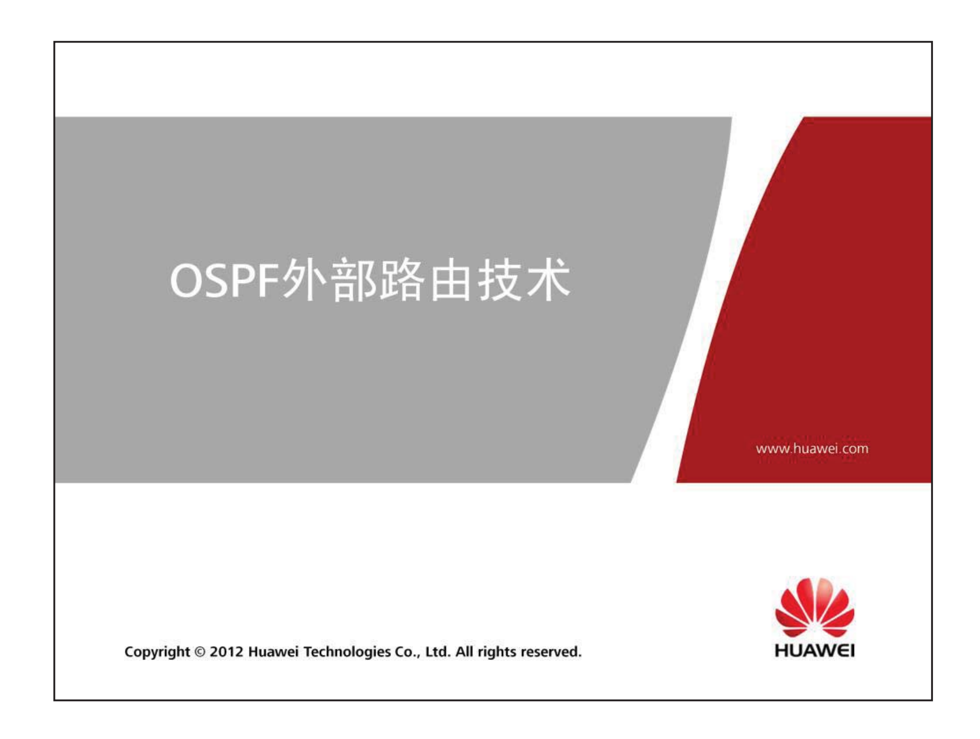
如何确定虚连接的对端IP地址?

通过计算对端路由器的最短生成树找到对端路由器在虚连接上的IP地址。

区域间路由汇聚功能在什么路由器上配置?

在区域边界路由器（ABR）上配置。


谢谢
www.huawei.com


OSPF外部路由技术
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍OSPF外部路由技术。

课程内容包括AS-External-LSA和ASBR-Summary-LSA的解释，外部路由类型，外部路由的Forwarding Address属性，配置外部路由引入和汇聚，配置OSPF多进程等内容。

Page1





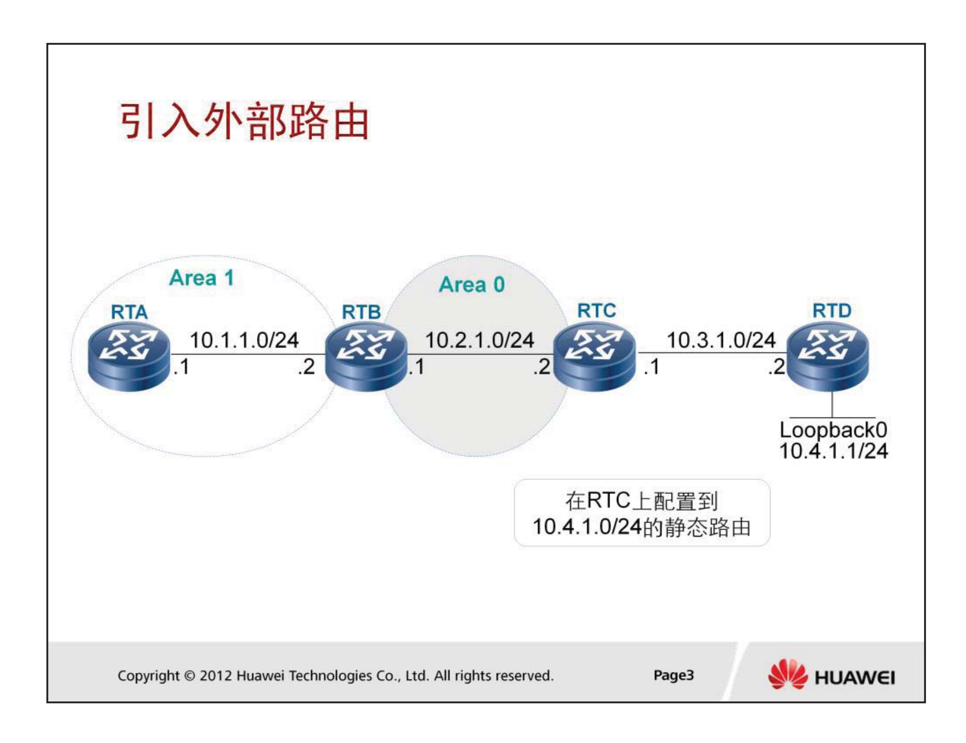
培训目标

学完本课程后，您应该能：

- 理解外部路由使用的LSA
- 理解外部路由类型
- 理解Forwarding Address属性
- 掌握外部路由引入的配置
- 掌握OSPF多进程的配置

Page2





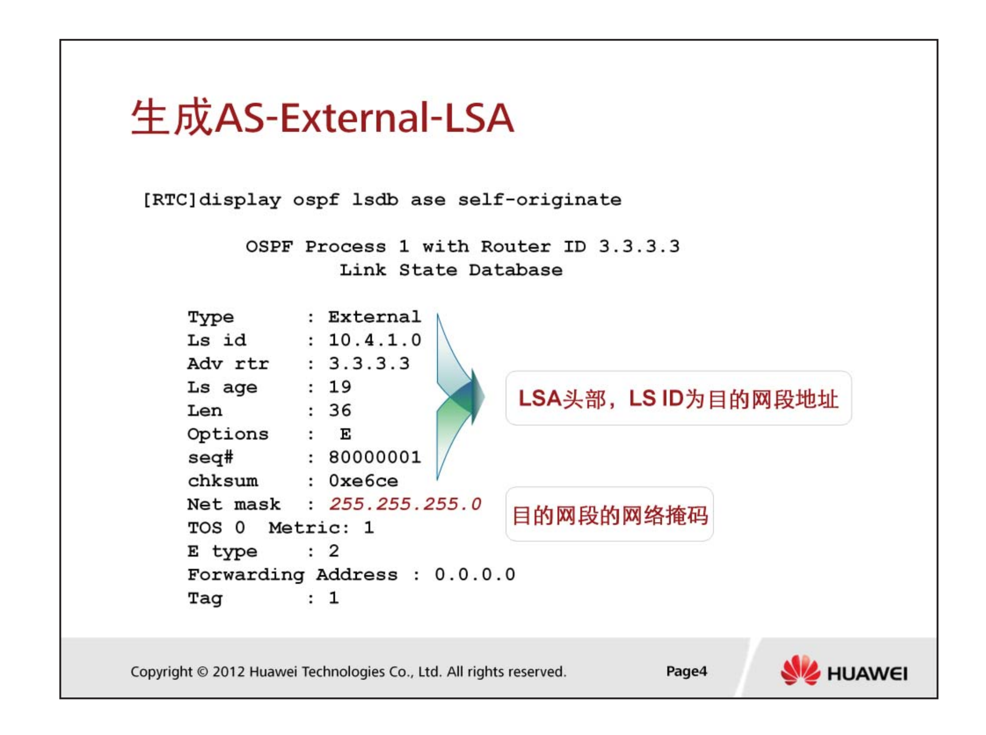
本例中：

在RTC上配置一条到10.4.1.0/24的静态路由，并将该静态路由做为外部路由引入OSPF。

因此，RTC是一个ASBR，RTB是一个ABR。

RTC会生成一条AS-External-LSA描述引入的外部路由，RTB会生成一条ASBR-Summary-LSA描述如何到达ASBR（RTC）。

AS-External-LSA用于描述如何从ASBR到达外部目的地；ASBR-Summary-LSA用于描述如何从ABR到达ASBR。



这是由RTC生成的AS-External-LSA。

AS-External-LSA中LSA头部信息设置如下：

Link State ID被设置为目的网段地址。

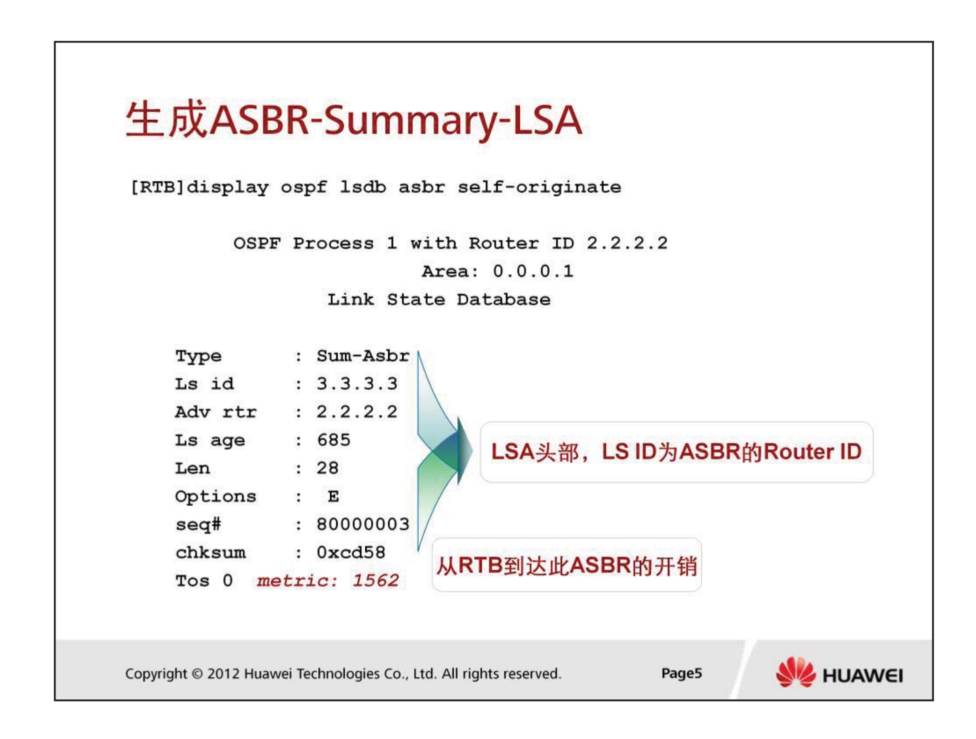
Advertising Router被设置为ASBR的Router ID。

其它字段设置如下：

Net mask被设置为目的网段的网络掩码。

Metric值可以在引入外部路由的时候指定，默认值为1。

外部路由信息可以携带一个Tag标签，用于传递该路由的附加信息，通常用于路由策略，默认值为1。



这是由RTB在Area 1内生成的ASBR-Summary-LSA。

ABR向区域外泛洪一条AS-External-LSA时，同时生成一条描述ASBR（该AS-External-LSA的Advertising Router）的ASBR-Summary-LSA向区域外泛洪。

在该ASBR-Summary-LSA中：

Link State ID被设置为该ASBR的Router ID；

Advertising Router被设置为该ABR的Router ID；

Metric设置为从该ABR到达此ASBR的OSPF开销。

第四类LSA只能在一个区域内泛洪，第五类LSA每泛洪到一个区域，相关的ABR都会生成一条新的第四类LSA来描述如何到达相关的ASBR，因此，描述到达同一个ASBR的第四类LSA可以有多条，其Advertising Router和metric是不同的，表示是由不同的ABR生成的。

含有AS-External-LSA的LSDB

```
[RTA]display ospf lsdb

        OSPF Process 1 with Router ID 1.1.1.1
                Link State Database

                        Area: 0.0.0.1
 Type       LinkState ID    AdvRouter         Age  Len    Sequence   Metric
 Router     2.2.2.2         2.2.2.2          1535  48     80000006    1562
 Router     1.1.1.1         1.1.1.1          1539  48     80000005    1562
 Sum-Net    10.2.1.0        2.2.2.2          1475  28     80000003    1562
 Sum-Asbr   3.3.3.3         2.2.2.2          1428  28     80000003    1562


                        AS External Database
 Type       LinkState ID    AdvRouter         Age  Len    Sequence   Metric
 External   10.4.1.0        3.3.3.3          1426  36     80000003       1
```

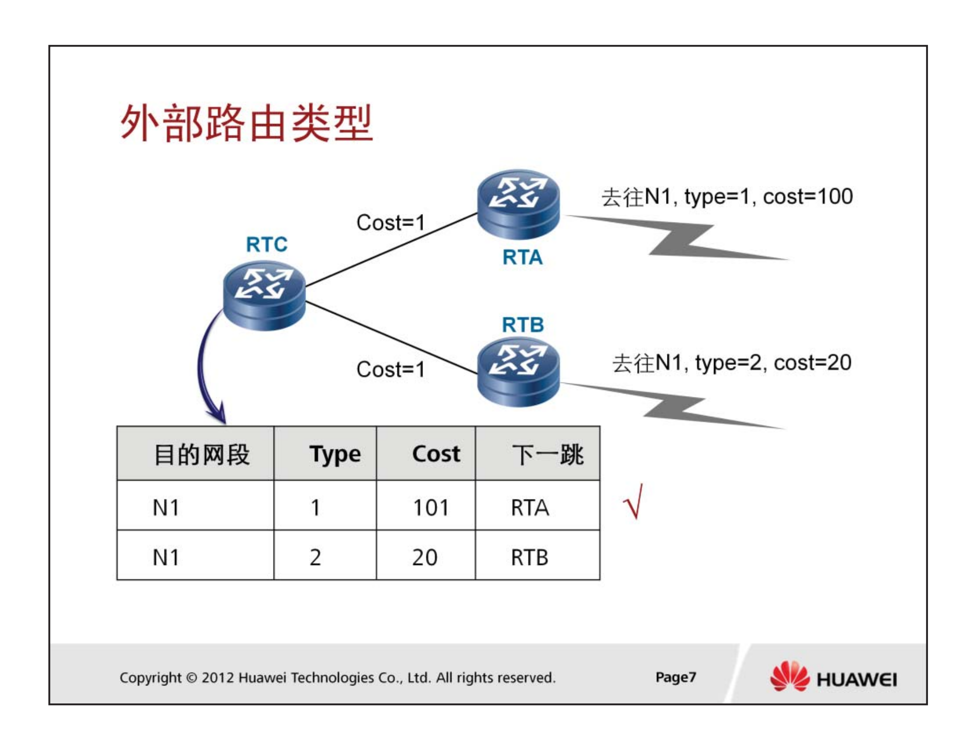
 Page6 

这是RTA的LSDB。

在该LSDB中，有一条ASBR-Summary-LSA和一条AS-External-LSA。

在LSDB中，外部路由与AS内部的路由的链路状态信息是分离的。

AS-External-LSA不属于任何区域。



OSPF共有两类外部路由:

第一类外部路由的AS外部开销值被认为和AS内部开销值是同一数量级的，因此第一类外部路由的开销值为AS内部开销值（路由器到ASBR的开销）与AS外部开销值之和;

第二类外部路由的AS外部开销值被认为远大于AS内部开销值，因此第二类外部路由的开销值只是AS外部开销值，忽略AS内部开销值。

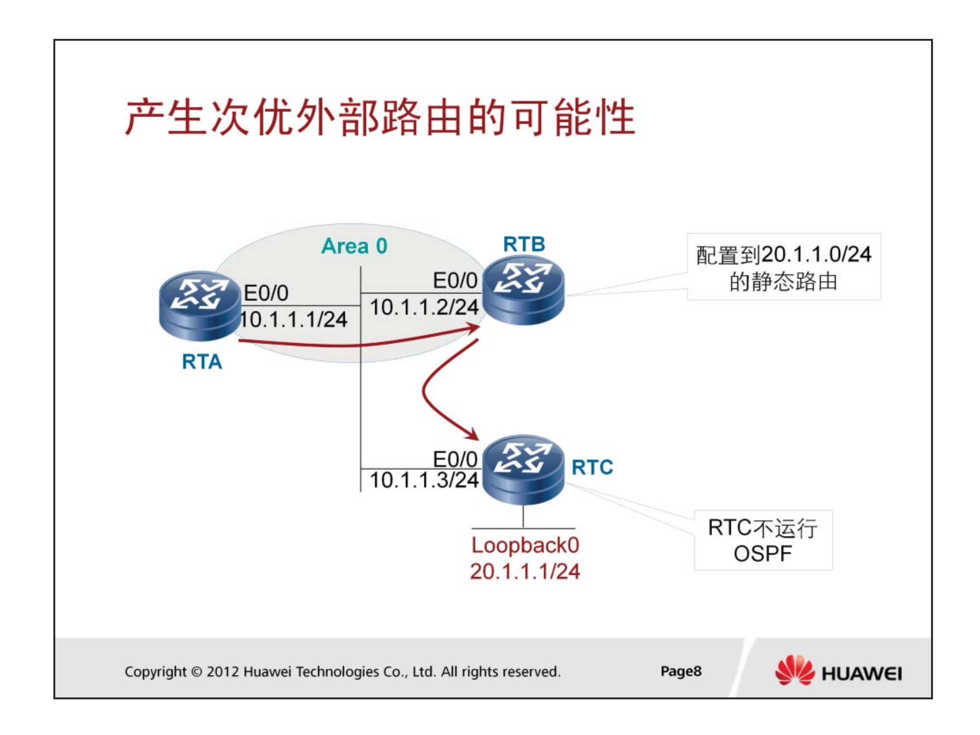
第一类外部路由永远比第二类外部路由优先。VRP中引入的外部路由类型缺省为第二类。

本例中:

RTA通告一条去往N1的AS外部路由，类型为1，开销为100。

RTB也通告一条去往N1的AS外部路由，类型为2，开销为20。

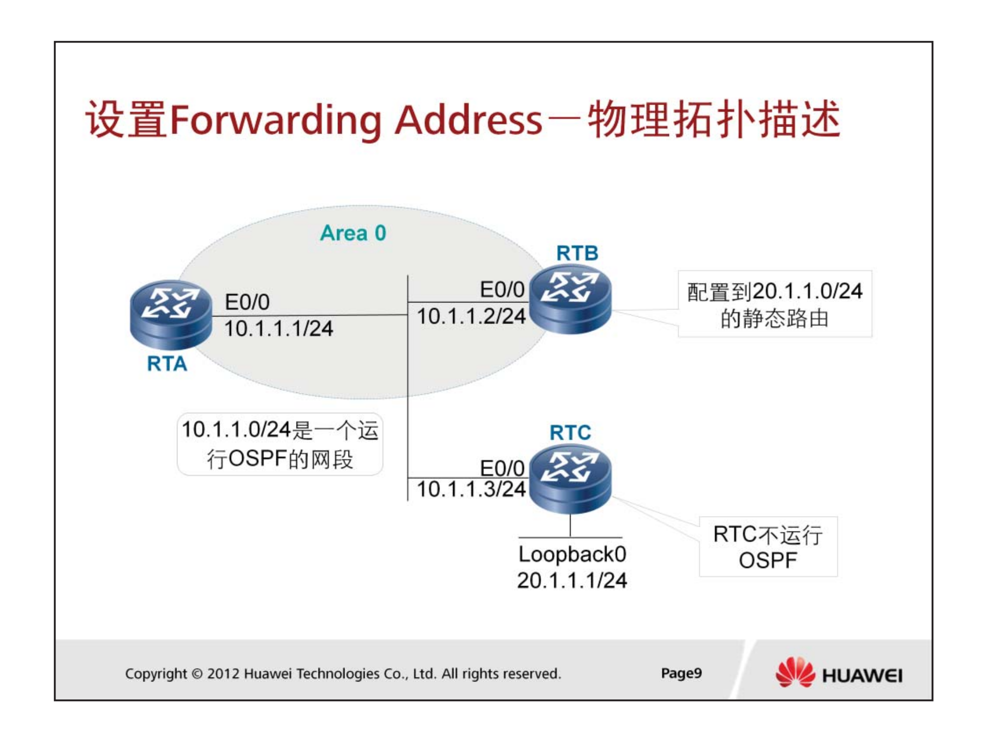
RTC收到RTA和RTB的Type5 LSA，由于RTA宣告的外部路由类型为Type1，所以RTC认为通过RTA去往N1的路由开销为100+1=101，RTB宣告的外部路由类型为Type2，所以RTC认为通过RTB去往N1的路由开销为20（忽略AS内部开销），由于第一类外部路由比第二类外部路由优先，所以RTC选择RTA做为去往N1的下一跳，尽管开销值看上去更大一些。



如图所示组网，10.1.1.0/24属于OSPF路由域；RTC不运行OSPF。

在RTB上配置到达RTC的Loopback0接口的静态路由，并做为外部路由引入到OSPF中，则RTA可以通过OSPF学习到这条外部路由，但是下一跳是RTB，因此在RTA上，这条路由是次优的，最优的下一跳应当为RTC的E0/0接口。

OSPF通过设置Forwarding Address来解决这个问题。



本例中：

在RTA和RTB上开启OSPF，将网段10.1.1.0/24配置在Area 0内。

RTA的Router ID设为1.1.1.1，RTB为2.2.2.2。

RTC不运行OSPF，网段10.1.1.0/24不在OSPF路由域内。

在RTB上，定义一个到RTC的Loopback0（20.1.1.0/24）的静态路由，并引入到OSPF中。

通过这个例子，探讨如何使用Forwarding Address实现最优路由的选择。

设置Forwarding Address

```
[RTB]ip route-static 20.1.1.0 24 10.1.1.3
[RTB]
[RTB]display ospf lsdb ase
         OSPF Process 1 with Router ID 2.2.2.2
                 Link State Database
    Type       : External
    Ls id      : 20.1.1.0
    Adv rtr    : 2.2.2.2
    Ls age     : 67
    Len        : 36
    Options    :  E
    seq#       : 80000001
    chksum     : 0x584b
    Net mask   : 255.255.255.0
    TOS 0  Metric: 1
    E type     : 2
    Forwarding Address : 10.1.1.3
    Tag        : 1
```

外部路由下一跳在OSPF路由域内

Forwarding Address被设置成10.1.1.3

 Page10 HUAWEI

通常情况下，生成AS-External-LSA时，Forwarding Address设置为0.0.0.0
。

但是如果引入到OSPF中的外部路由的下一跳在一个OSPF路由域内，则在描述该外部路由的AS-External-LSA中，Forwarding Address应当被设置为ASBR路由表中该路由的下一跳。

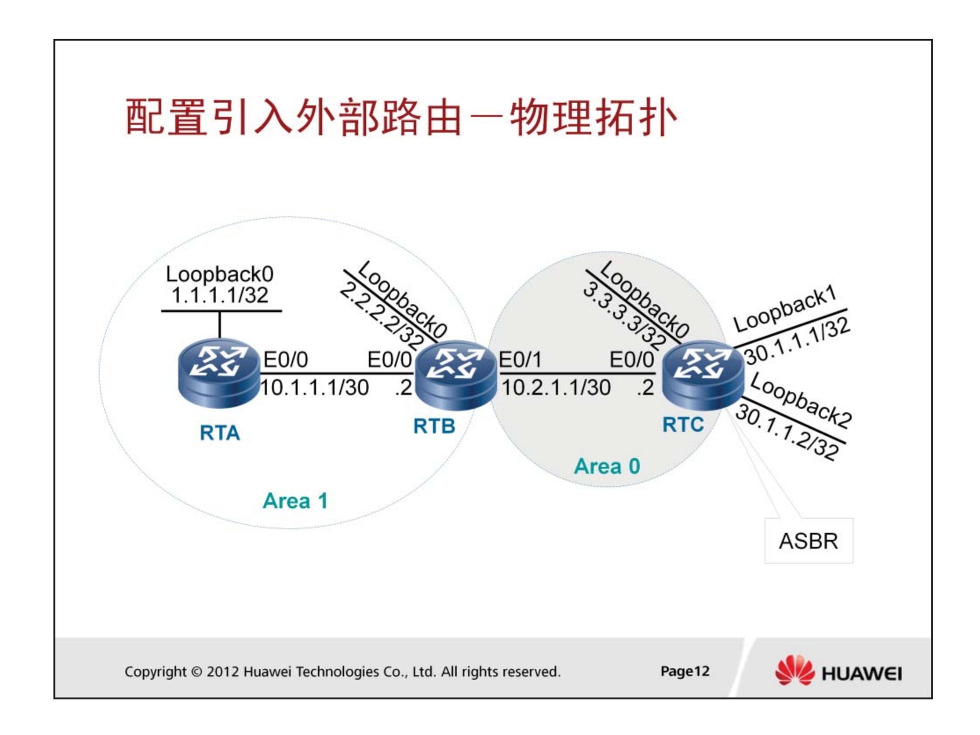
本例中，RTB定义该静态路由的下一跳为10.1.1.3，在OSPF路由域内，因此在RTB生成的AS-External-LSA中， Forwarding Address被设置成10.1.1.3。



RTA收到此LSA以后，计算AS外部路由。

如果Forwarding Address没有被设置为0，则该路由的下一跳就是路由表中到Forwarding Address的下一跳。

本例中，Forwarding Address被设置为10.1.1.3，所以通过计算，该外部路由的下一跳仍为10.1.1.3，和到Forwarding Address的下一跳是一致的（此例中， Forwarding Address 和到Forwarding Address的下一跳相同）。



本例中：

RTC为ASBR。

在RTC上配置OSPF引入直连路由，因此，30.1.1.1/32和30.1.1.2/32两个网段会被做为外部路由引入到OSPF中；

在RTC上配置路由汇聚，把路由条目30.1.1.1/32和30.1.1.2/32汇聚成30.1.1.0/24。配置完成后，RTC将向外通告一条到达30.1.1.0/24的路由条目，不通告到达30.1.1.1/32和30.1.1.2/32明细路由条目。

配置引入外部路由－配置RTA和RTB

```
[RTA]router id 1.1.1.1
[RTA]ospf
[RTA-ospf-1]area 1
[RTA-ospf-1-area-0.0.0.1]network 1.1.1.1 0.0.0.0
[RTA-ospf-1-area-0.0.0.1]network 10.1.1.0 0.0.0.3
[RTA-ospf-1-area-0.0.0.1]return
<RTA>

[RTB]router id 2.2.2.2
[RTB]ospf
[RTB-ospf-1]area 1
[RTB-ospf-1-area-0.0.0.1]network 2.2.2.2 0.0.0.0
[RTB-ospf-1-area-0.0.0.1]network 10.1.1.0 0.0.0.3
[RTB-ospf-1-area-0.0.0.1]quit
[RTB-ospf-1]area 0
[RTB-ospf-1-area-0.0.0.0]network 10.2.1.0 0.0.0.3
[RTB-ospf-1-area-0.0.0.0]return
<RTB>
```

　Page13　HUAWEI

在RTA上，两个网段配置在Area 1中。

在RTB上，网段2.2.2.2/32和10.1.1.0/30在Area 1中，网段10.2.1.0/30在Area 0中。

RTB为Area 1的ABR，因此RTB在向Area 1中通告AS-External-LSA时会同时生成一条ASBR-Summary-LSA描述ASBR（RTC）。

配置引入外部路由－配置RTC

```
[RTC]router id 3.3.3.3
[RTC]ospf
[RTC-ospf-1]area 0
[RTC-ospf-1-area-0.0.0.0]network 3.3.3.3 0.0.0.0
[RTC-ospf-1-area-0.0.0.0]network 10.2.1.0 0.0.0.3
[RTC-ospf-1-area-0.0.0.0]quit
[RTC-ospf-1]
[RTC-ospf-1]import-route direct
[RTC-ospf-1]
[RTC-ospf-1]asbr-summary 30.1.1.0 255.255.255.0
[RTC-ospf-1]quit
[RTC]
```

把直连路由引入OSPF

在ASBR上执行路由汇聚

 Page14 HUAWEI

在RTC上，引入直连路由，然后配置外部路由汇聚。

引入外部路由：

import-route { limit [limit-number] | protocol [cost value] [type value] [tag value] [route-policy route-policy-name] }

引入外部路由时，可以配置路由开销，外部路由类型和tag值。

limit：指定一个OSPF进程中可引入的最大外部路由数量。

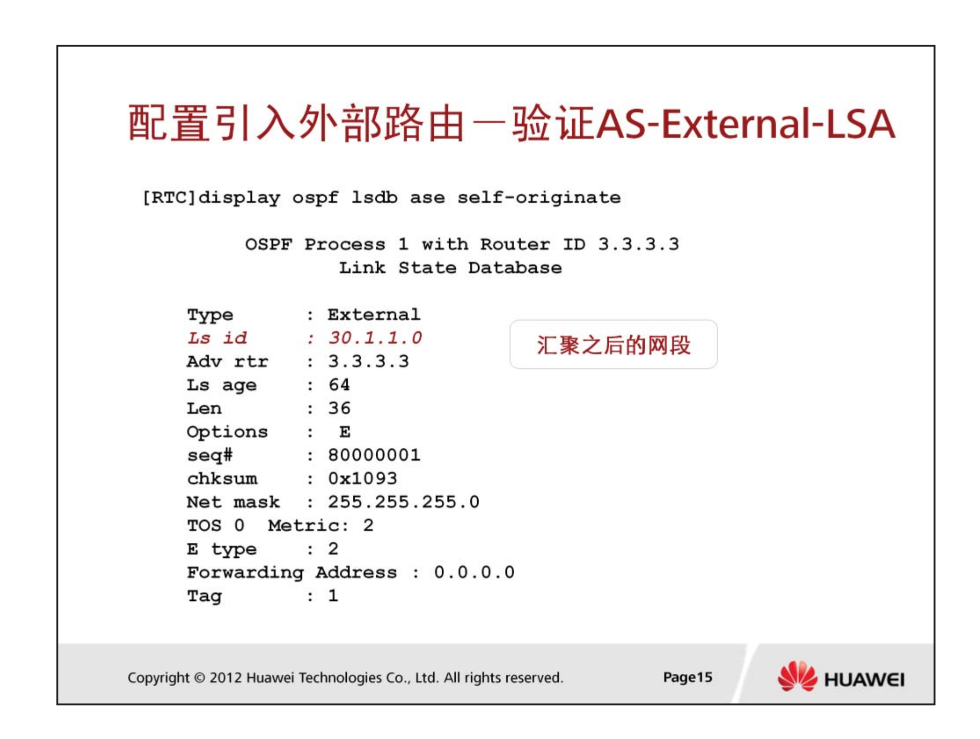
外部路由汇聚：

asbr-summary：ip-address mask [not-advertise | tag tag-value]

not-advertise：不通告匹配指定IP地址/掩码的路由。如果不指定该参数将通告聚合路由。

tag-value：用于通过Route-policy控制路由发布，tag-value的取值范围为0～4294967295。如果不指定该参数，缺省值为1。

缺省情况下，只通告聚合之后的路由。



在RTC产生的第五类LSA中，只有描述到达网段30.1.1.0/24的LSA，没有描述到达明细网段30.1.1.1/32和30.1.1.2/32的LSA。

配置引入外部路由－验证ASBR-Summary-LSA

```
[RTB]display ospf lsdb asbr self-originate

         OSPF Process 1 with Router ID 2.2.2.2
                         Area: 0.0.0.0
                 Link State Database

                         Area: 0.0.0.1
                 Link State Database

    Type        : Sum-Asbr
    Ls id       : 3.3.3.3
    Adv rtr     : 2.2.2.2
    Ls age      : 264
    Len         : 28
    Options     :  E
    seq#        : 80000001
    chksum      : 0xa0a6
    Tos 0  metric: 1
```

由RTB产生，描述RTC

Page16 HUAWEI

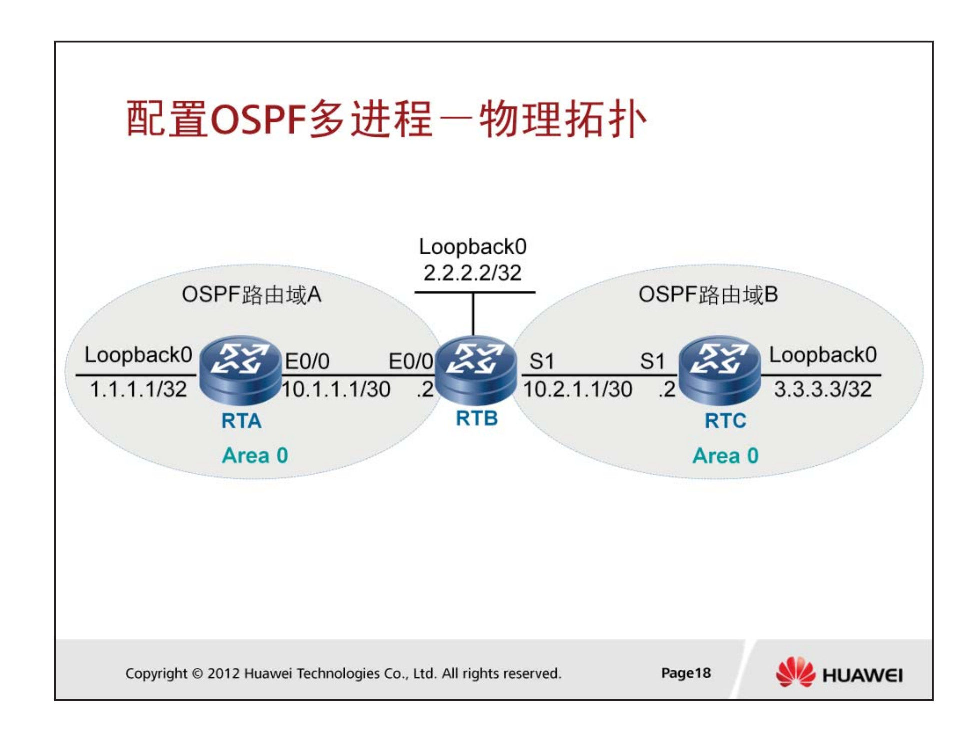
在RTB上会产生一条第四类LSA，描述如何从RTB到达ASBR（RTC），并在Area 1中泛洪。

配置引入外部路由—验证RTA的路由表

```
[RTA]display ip routing-table
Routing Tables: Public
         Destinations : 17        Routes : 17
Destination/Mask    Proto  Pre  Cost       NextHop         Interface
        1.1.1.1/32  Direct 0    0          127.0.0.1       InLoopBack0
        3.3.3.3/32  O_ASE  150  1          10.1.1.2        Ethernet0
       10.1.1.0/24  O_ASE  150  1          10.1.1.2        Ethernet0
       10.1.1.0/30  Direct 0    0          10.1.1.1        Ethernet0
       10.1.1.1/32  Direct 0    0          127.0.0.1       InLoopBack0
       10.2.1.0/30  OSPF   10   2          10.1.1.2        Ethernet0
       10.2.1.1/32  O_ASE  150  1          10.1.1.2        Ethernet0
       10.3.1.0/24  O_ASE  150  1          10.1.1.2        Ethernet0
       10.3.1.2/32  O_ASE  150  1          10.1.1.2        Ethernet0
       10.5.1.0/30  Direct 0    0          10.5.1.1        Serial1
       10.5.1.1/32  Direct 0    0          127.0.0.1       InLoopBack0
       10.5.1.2/32  Direct 0    0          10.5.1.2        Serial1
       20.1.1.1/32  Direct 0    0          127.0.0.1       InLoopBack0
       20.1.1.2/32  Direct 0    0          127.0.0.1       InLoopBack0
       30.1.1.0/24  O_ASE  150  2          10.1.1.2        Ethernet0
      127.0.0.0/8   Direct 0    0          127.0.0.1       InLoopBack0
      127.0.0.1/32  Direct 0    0          127.0.0.1       InLoopBack0
```

 Page17 HUAWEI

在AS内部路由器的路由表中，只有汇聚之后的到达网段30.1.1.0/24的路由条目，没有明细路由条目。



本例中：

网络中有两个OSPF路由域。

做为ASBR，在RTB上应启用两个OSPF进程。

RTA启用OSPF进程1，RTC也启用OSPF进程1。

注意：相邻的两台路由器启用不同的OSPF进程也可以正常通信，因为在OSPF报文中不含有进程信息，只要其他参数协商通过就可以正常通信。

配置OSPF多进程－配置RTA

```
[RTA]router id 1.1.1.1
[RTA]ospf
[RTA-ospf-1]area 0
[RTA-ospf-1-area-0.0.0.0]network 1.1.1.1 0.0.0.0
[RTA-ospf-1-area-0.0.0.0]network 10.1.1.0 0.0.0.3
[RTA-ospf-1-area-0.0.0.0]quit
[RTA-ospf-1]quit
[RTA]
```

 Page19 

在RTA上启用一个OSPF进程，将所有网段配置在Area 0中。

配置OSPF多进程—配置RTB

```
[RTB]ospf 1 router-id 2.2.2.2
[RTB-ospf-1]area 0
[RTB-ospf-1-area-0.0.0.0]network 10.1.1.0 0.0.0.3
[RTB-ospf-1-area-0.0.0.0]quit
[RTB-ospf-1]import-route direct
[RTB-ospf-1]import-route ospf 2
[RTB-ospf-1]quit
[RTB]ospf 2 router-id 2.2.2.2
[RTB-ospf-2]area 0
[RTB-ospf-2-area-0.0.0.0]network 10.2.1.0 0.0.0.3
[RTB-ospf-2-area-0.0.0.0]quit
[RTB-ospf-2]import-route direct
[RTB-ospf-2]import-route ospf 1
[RTB-ospf-2]quit
[RTB]
```

可以为不同的进程指定不同的Router ID

把OSPF进程2的路由引入到OSPF进程1中

把OSPF进程1的路由引入到OSPF进程2中

 Page20 HUAWEI

在RTB上启用两个OSPF进程，为每个OSPF进程分别指定Router ID。

配置每个进程时，把另外一个进程的路由信息引入到本进程中。

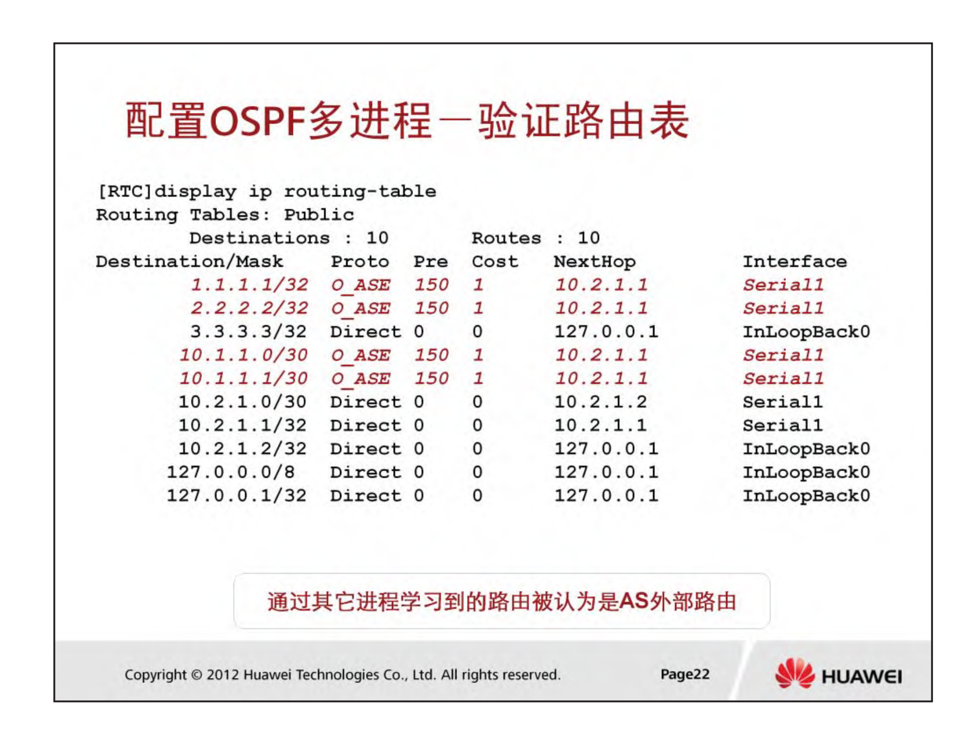
配置OSPF多进程－配置RTC

```
[RTC]router id 3.3.3.3
[RTC]ospf 2
[RTC-ospf-2]area 0
[RTC-ospf-2-area-0.0.0.0]network 3.3.3.3 0.0.0.0
[RTC-ospf-2-area-0.0.0.0]network 10.2.1.0 0.0.0.3
[RTC-ospf-2-area-0.0.0.0]quit
[RTC-ospf-2]quit
[RTC]
```

 Page21



在RTC上开启一个OSPF进程，把所有网段配置在Area 0中。



从其它OSPF进程中学习到的路由条目均被认为是AS外部路由。

配置OSPF多进程—验证LSDB

```
[RTB]display ospf lsdb
        OSPF Process 1 with Router ID 2.2.2.2
                Link State Database
                        Area: 0.0.0.0
 Type      LinkState ID    AdvRouter  Age  Len   Sequence   Metric
 Router    2.2.2.2         2.2.2.2    136  36    80000006        1
 Router    1.1.1.1         1.1.1.1    469  36    80000003        1
 Network   10.1.1.2        2.2.2.2    463  32    80000002        0
                AS External Database
 Type      LinkState ID    AdvRouter  Age  Len   Sequence   Metric
 External  10.2.1.0        2.2.2.2    136  36    80000001        1

        OSPF Process 2 with Router ID 2.2.2.2
                Link State Database
                        Area: 0.0.0.0
 Type      LinkState ID    AdvRouter  Age  Len   Sequence   Metric
 Router    3.3.3.3         3.3.3.3    200  48    80000024        1
 Router    2.2.2.2         2.2.2.2    120  48    80000022        1
                AS External Database
 Type      LinkState ID    AdvRouter  Age  Len   Sequence   Metric
 External  10.1.1.0        2.2.2.2    120  36    80000001        1
```

 Page23 HUAWEI

RTB上有两个LSDB，每个进程拥有独立的LSDB。



问 题

第四类和第五类LSA分别在什么路由器上产生?

两种类型的外部路由有什么区别?

Forwarding Address字段如何设置?

 Page24



第四类和第五类LSA分别在什么路由器上产生?

第四类LSA在ABR上产生，每条第四类LSA只在一个区域内泛洪；第五类LSA在ASBR上产生，在整个AS内泛洪。

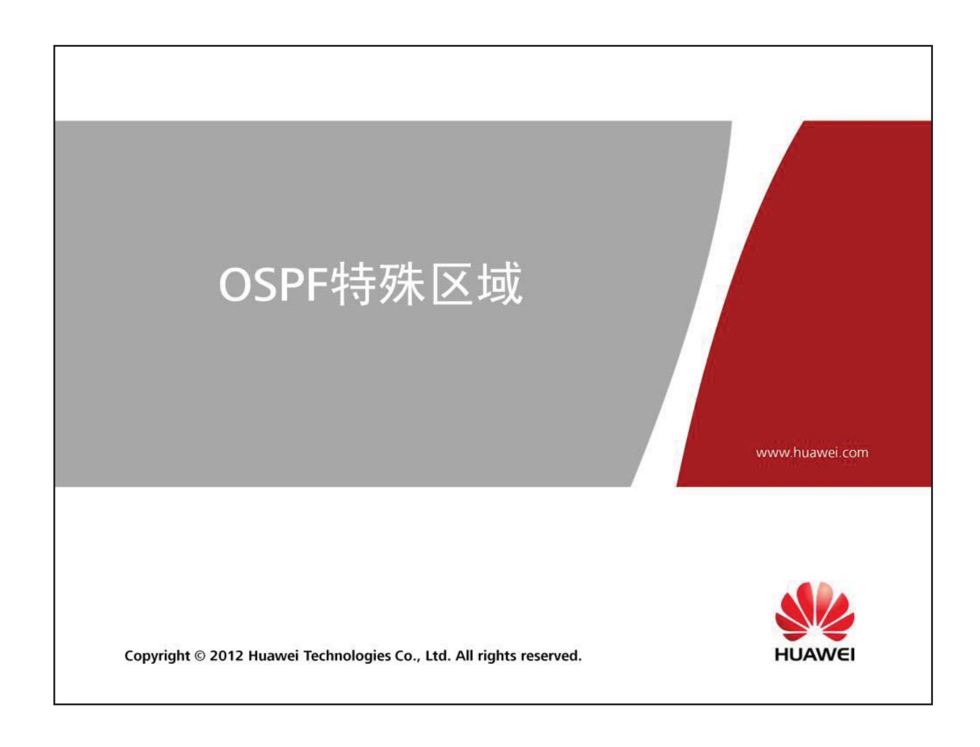
两种类型的外部路由有什么区别?

开销值计算不同，第一类外部路由的AS外部开销和AS内部开销是相同数量级的；第二类外部路由的AS外部开销被认为远大于AS内部开销。

Forwarding Address字段如何设置?

在ASBR上，所引入外部路由的下一跳如果在OSPF路由域内，则Forwarding Address应设置为此外部路由的下一跳，如果所引入外部路由的下一跳不在该OSPF路由域内，则Forwarding Address应设置为0.0.0.0。


谢谢
www.huawei.com


OSPF特殊区域
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍OSPF特殊区域。

课程内容包括Stub区域，完全Stub区域，NSSA等内容。





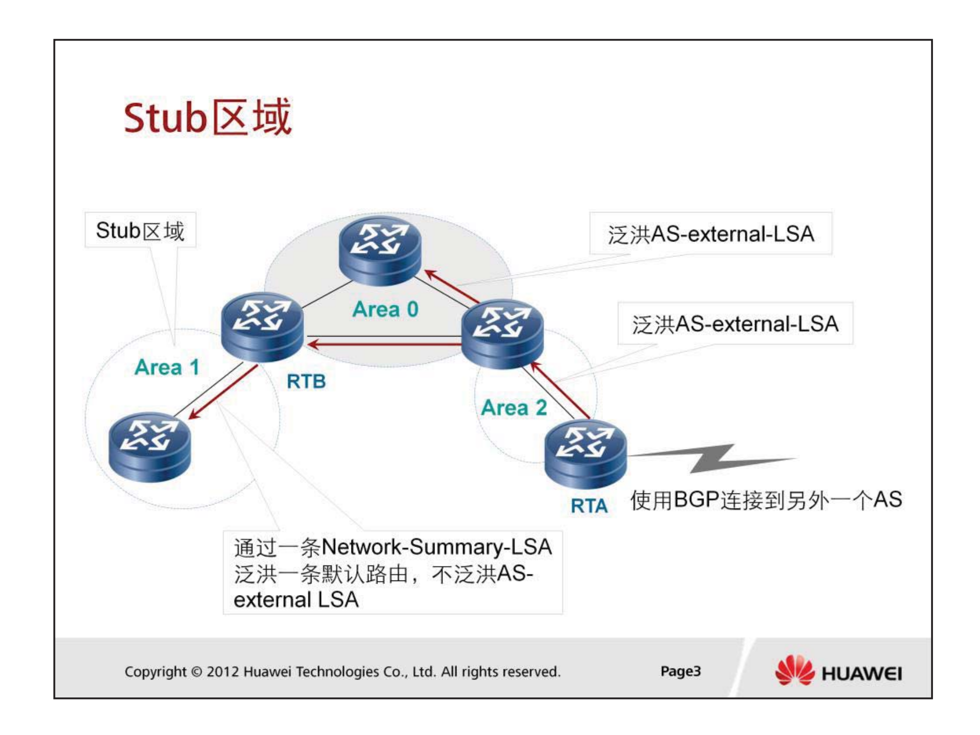
培训目标

学完本课程后，您应该能：

- 理解Stub区域的概念和配置
- 理解完全Stub区域的概念和配置
- 理解NSSA的概念和配置

Page2





OSPF允许将特定区域配置为Stub区域。

AS-external-LSA不允许被发布到Stub区域内。到AS外部的路由只能基于由ABR生成的一条默认路由。

Stub区域技术可以减少Stub区域内部路由器上LSDB的规模和对内存的需求。

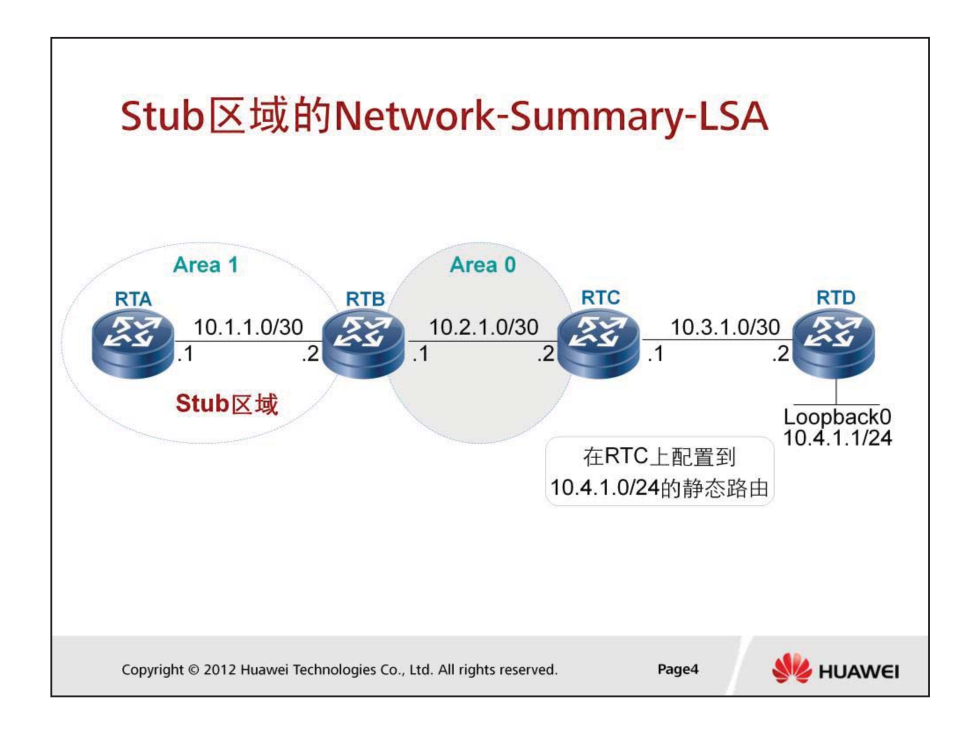
虚连接不能跨越Stub area。

本例中：

RTA是一个ASBR，将Area 1配置为Stub area，RTB是Area 1的ABR。

RTA将AS外部的路由信息通过AS-external-LSA向AS内部泛洪。

RTB只通过一条Network-Summary-LSA向Area 1内通告一条默认路由，不泛洪任何AS-external-LSA。

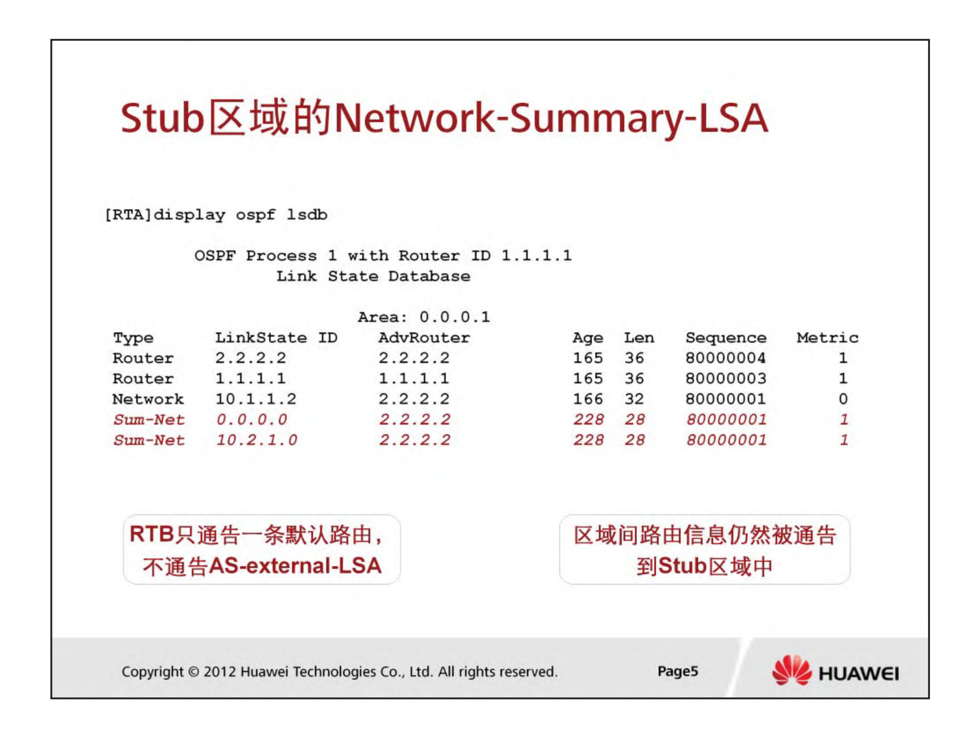


本例中：

在RTC上配置一条到达10.4.1.0/24的静态路由，并将静态路由引入到OSPF中。因此，RTC是一个ASBR。

RTA使用1.1.1.1做为Router ID；RTB使用2.2.2.2做为Router ID；RTC使用3.3.3.3做为Router ID。

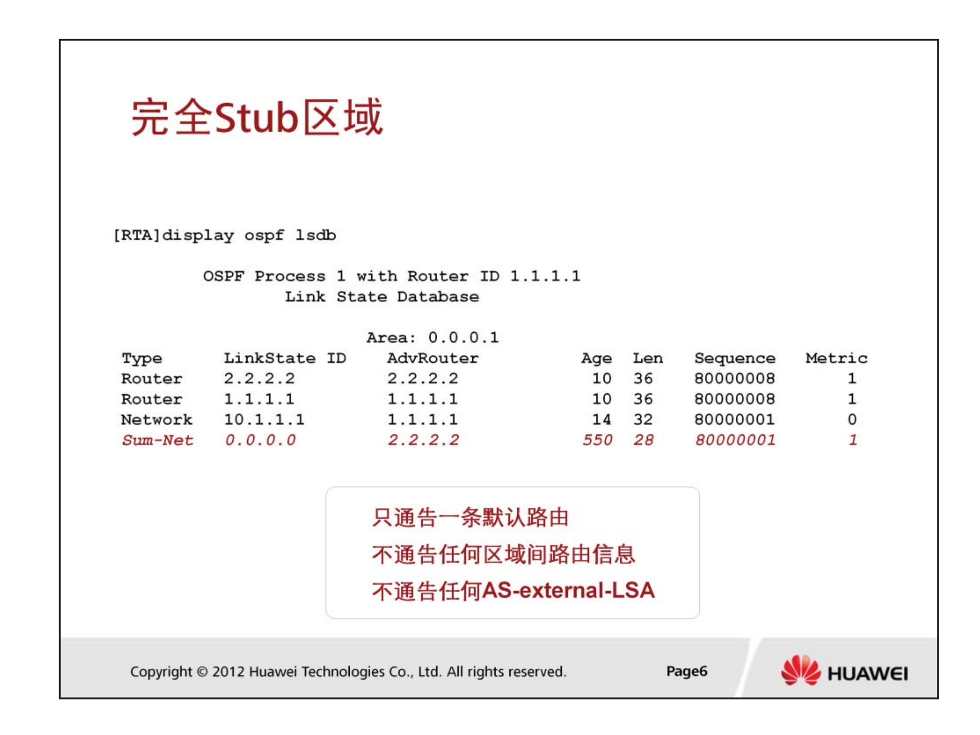
将Area 1配置成Stub区域。



这是由RTB在Area 1内产生的Network-Summary-LSA。

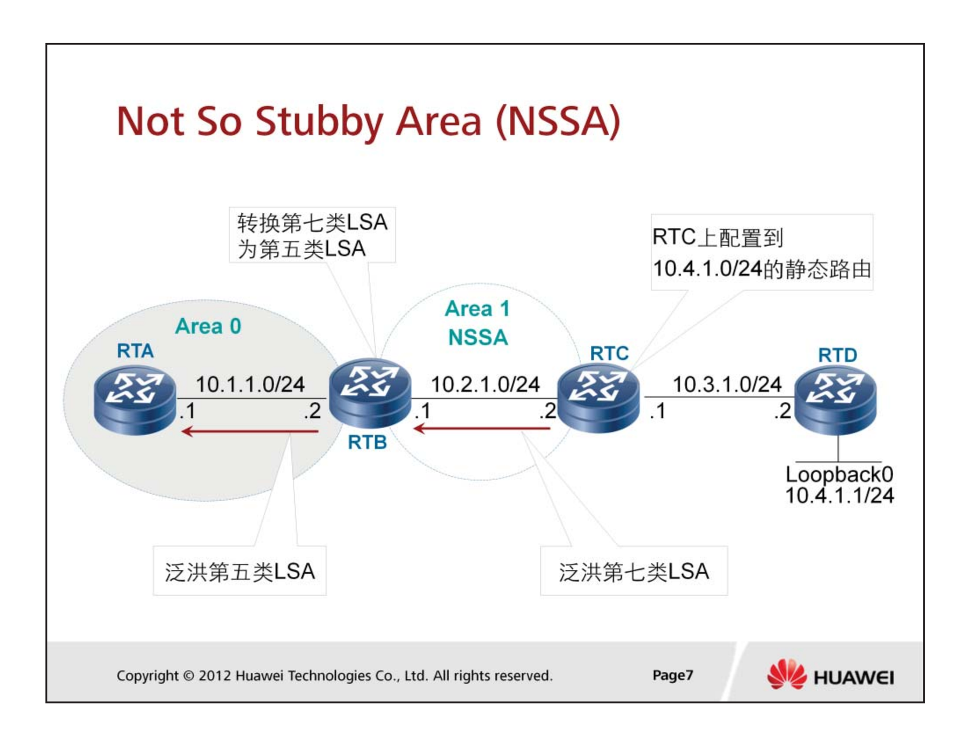
RTB通过一条Network-Summary-LSA描述一条默认路由，并在Area 1内泛洪。

区域间的路由信息仍然需要发布到Stub区域中。



通过配置ABR可以开启一个Stub区域的完全Stub功能。

如果一个区域被配置为完全Stub区域，只有一条由Network-Summary-LSA描述的默认路由被通告到该区域中，没有区域间的路由信息和AS外部的路由信息。即完全Stub区域的ABR不向该区域中泛洪Summary-LSA（除了默认路由）和AS-external-LSA。



NSSA是指Not So Stubby Area，表示不那么 “Stub”的区域。

NSSA和Stub区域有些类似，都不能处理AS-external-LSA，但是NSSA可以用另外一种方式引入外部路由。

本例中：

在RTA、RTB、RTC上分别创建Loopback0接口，配置IP地址分别为1.1.1.1/32、2.2.2.2/32、3.3.3.3/32。

RTA的Router ID为1.1.1.1，RTB为2.2.2.2，RTC为3.3.3.3。

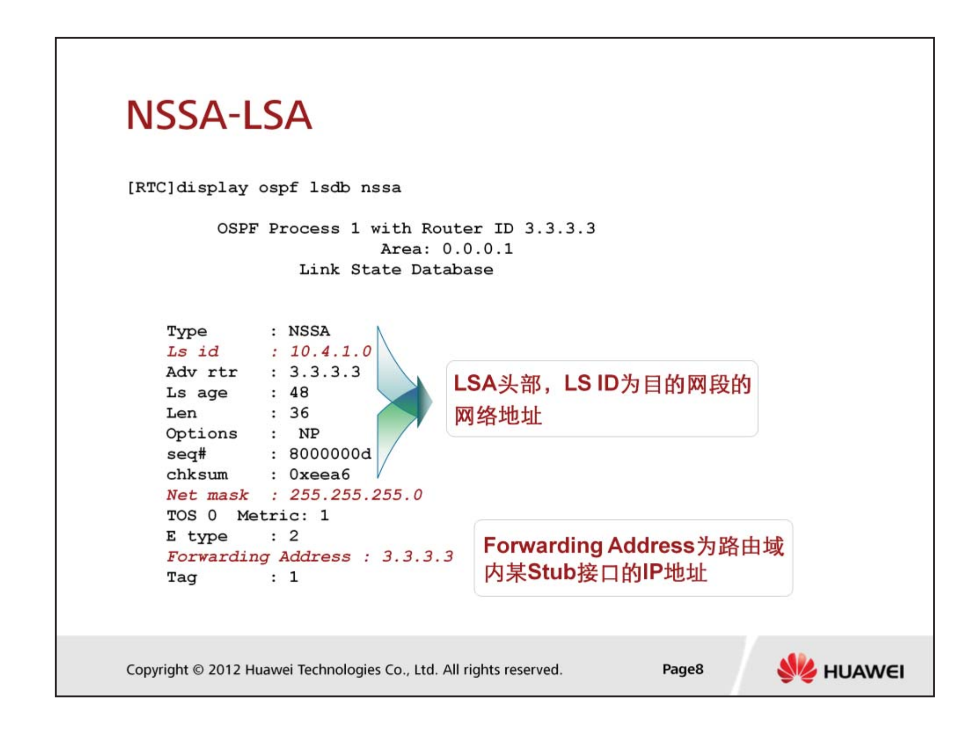
配置网段1.1.1.1/32、10.1.1.0/24在Area 0内；配置网段2.2.2.2/32、10.2.1.0/24和网段3.3.3.3/32在Area 1内。

在RTC上配置一条到达10.4.1.0/24的静态路由，下一跳为10.3.1.2，将静态路由引入到OSPF中。

将Area 1配置为NSSA。

RTC将泛洪一条NSSA-LSA（第七类LSA）到Area 1中，该LSA用于描述该外部路由。

第五类LSA不在NSSA内泛洪，第七类LSA不在NSSA外泛洪，因此NSSA的ABR（RTB）需要将该NSSA-LSA转换成一条AS-external-LSA，在AS内的其他区域内泛洪。



这是由RTC生成的NSSA-LSA（第七类LSA）。

本例中，在RTC上引入静态路由到OSPF后，RTC生成一条NSSA-LSA，描述一条到达AS外部网段10.4.1.0/24的路由信息，重要字段的设置规则如下：

LS ID设置为目的网段的网络地址；

Options字段显示这条LSA可以被ABR转换成一条AS-external-LSA；

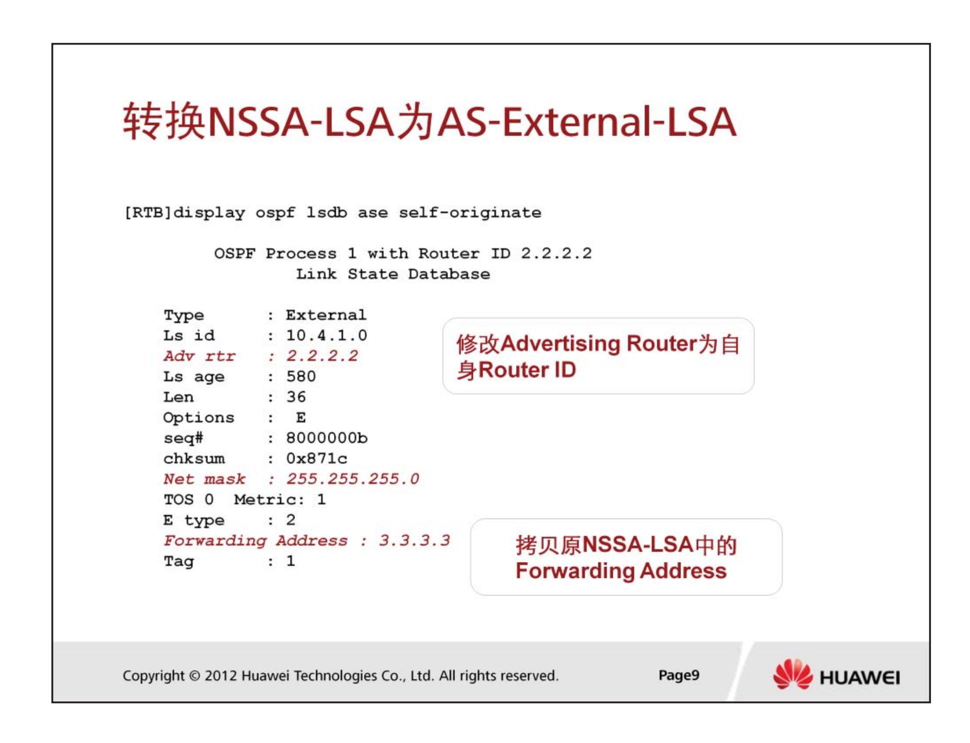
Metric默认设置为1；

E type（外部路由类型）默认设置为第二类；

另外，NSSA-LSA的Forwarding Address设置规则如下：

如果Options字段显示此LSA不可以被转换成第五类LSA，则Forwarding Address可以被设置成0.0.0.0；

如果Options字段显示此LSA可以被转换成第五类LSA，则Forwarding Address不能被设置成0.0.0.0。如果所引入外部路由的下一跳在OSPF路由域内，则Forwarding Address直接设置为所引入外部路由的下一跳；如果所引入外部路由的下一跳不在OSPF路由域内，则Forwarding Address设置为该ASBR上某个OSPF路由域内的Stub网段（例如Loopback0接口）的接口IP地址，有多个Stub网段时选IP地址最大者。



当RTB（Area 1的ABR）收到这条NSSA-LSA以后，将这条LSA转换成一条AS-external-LSA，重要字段处理如下：

LS ID和Network Mask两字段的值直接从原来NSSA-LSA中拷贝，描述到达AS外部网段10.4.1.0/24的路由信息；

不重新计算Metric值，如果配置时未指定Metric值，默认设置为1；

Forwarding Address直接从原NSSA-LSA中拷贝，不做修改。

因此，NSSA之外的路由器只根据所接收的AS-external-LSA中的Forwarding Address计算下一跳，并不计算如何到达ASBR。

为了防止外部路由信息重复，在一个NSSA有多个ABR的时候，只允许一个ABR可以把NSSA-LSA转换成AS-external-LSA，这个ABR称为此NSSA的Translator。

Translator基于Router ID选举。NSSA的ABR会在Router-LSA中使用一个Bit标识自己是NSSA的ABR，通过检查区域中的Router-LSA，每个NSSA的ABR都可以维护一个ABR列表，从中选举Router ID最大的做为Translator。

使用NSSA-LSA发布默认路由

```
[RTB]display ospf lsdb nssa self-originate

         OSPF Process 1 with Router ID 2.2.2.2
                         Area: 0.0.0.0
                 Link State Database

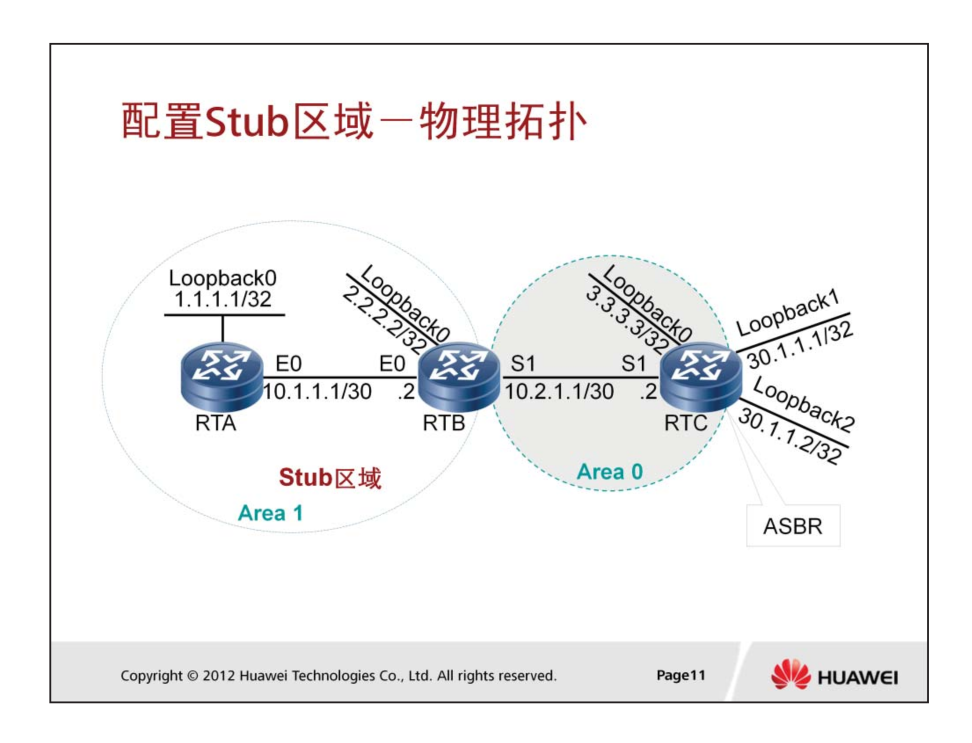
                         Area: 0.0.0.1
                 Link State Database

    Type       : NSSA
    Ls id      : 0.0.0.0
    Adv rtr    : 2.2.2.2
    Ls age     : 1231
    Len        : 36
    Options    : None
    seq#       : 80000001
    chksum     : 0x6654
    Net mask   : 0.0.0.0
    TOS 0  Metric: 1
    E type     : 2
    Forwarding Address : 2.2.2.2
    Tag        : 1
```

 Page10 HUAWEI

可以配置NSSA区域的ABR向NSSA区域内通告一条默认路由，但是这条默认路由是由一条NSSA-LSA描述，而不是由Network-Summary-LSA描述。

本例中，由RTB生成一条LS ID为0.0.0.0，Net mask为0.0.0.0的默认路由，并泛洪到Area 1内。



本例中:

在RTC上引入直连路由到OSPF，因此RTC是ASBR，网段30.1.1.1/32和网段30.1.1.2/32被做为外部路由引入。

配置Area 1为Stub区域，因此RTB不会向Area 1内泛洪第五类LSA，只会向区域内通告默认路由。

配置Stub区域－配置RTA和RTB

```
[RTA]router id 1.1.1.1
[RTA]ospf
[RTA-ospf-1]area 1
[RTA-ospf-1-area-0.0.0.1]network 1.1.1.1 0.0.0.0
[RTA-ospf-1-area-0.0.0.1]network 10.1.1.0 0.0.0.3
[RTA-ospf-1-area-0.0.0.1]stub
[RTA-ospf-1-area-0.0.0.1]return
<RTA>
```

指定Area 1为Stub区域

```
[RTB]router id 2.2.2.2
[RTB]ospf
[RTB-ospf-1]area 1
[RTB-ospf-1-area-0.0.0.1]network 2.2.2.2 0.0.0.0
[RTB-ospf-1-area-0.0.0.1]network 10.1.1.0 0.0.0.3
[RTB-ospf-1-area-0.0.0.1]stub
[RTB-ospf-1-area-0.0.0.1]quit
[RTB-ospf-1]area 0
[RTB-ospf-1-area-0.0.0.0]network 10.2.1.0 0.0.0.3
[RTB-ospf-1-area-0.0.0.0]return
<RTB>
```

 Page12 HUAWEI

必须在Stub区域中的每一台路由器上指定该区域为Stub区域。

在RTA和RTB上，把Area 1配置为Stub区域。

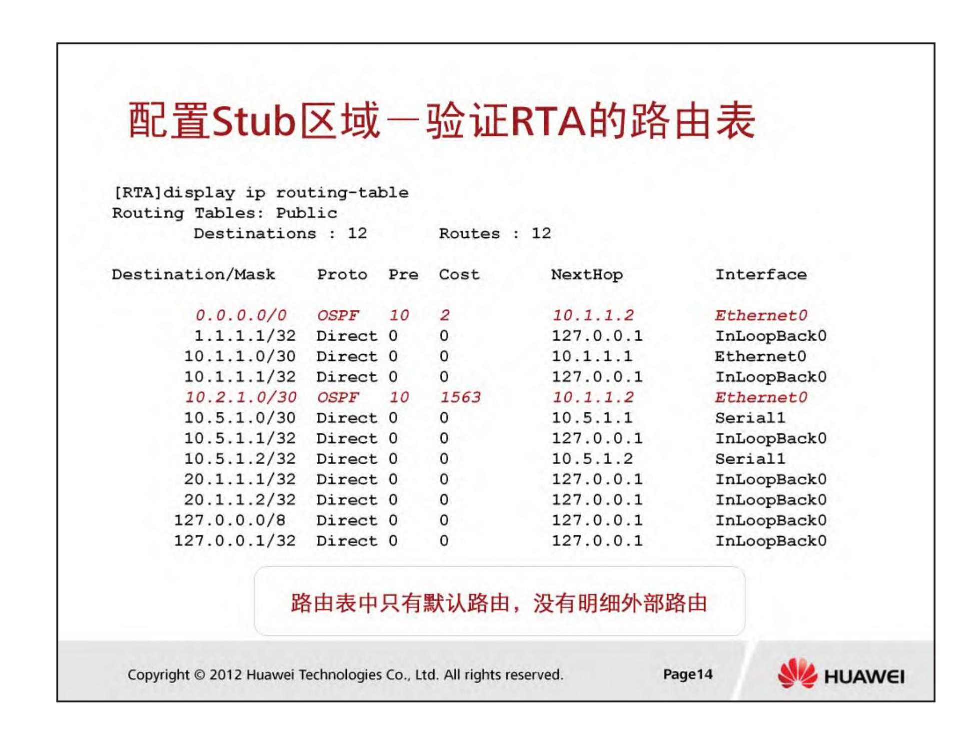
配置Stub区域－配置RTC

```
[RTC]router id 3.3.3.3
[RTC]ospf
[RTC-ospf-1]area 0
[RTC-ospf-1-area-0.0.0.0]network 3.3.3.3 0.0.0.0
[RTC-ospf-1-area-0.0.0.0]network 10.2.1.0 0.0.0.3
[RTC-ospf-1-area-0.0.0.0]quit
[RTC-ospf-1]import-route direct
[RTC-ospf-1]quit
[RTC]
```

引入直连路由

 Page13 HUAWEI

在RTC上，引入直连路由。



在RTA的路由表中，通过OSPF学习到的只有ABR通告的一条默认路由和AS内部路由，没有AS外部的明细路由。

配置完全Stub区域

```
[RTB]ospf
[RTB-ospf-1]area 1
[RTB-ospf-1-area-0.0.0.1]stub no-summary
[RTB-ospf-1-area-0.0.0.1]quit
[RTB-ospf-1]quit
[RTB]
```

No-summary是指不向区域内转发Summary-LSAs（第三类和第四类）

Page15


配置Stub区域后，通过在ABR（RTB）上执行命令stub no-summary，将Area 1配置为完全Stub区域。

配置一个区域为完全Stub区域以后，该区域的ABR不在该区域内产生、转发第三类、第四类LSA和第五类LSA，所以不会有区域间路由和AS外部路由被通告到完全Stub区域内。

配置完全Stub区域－验证RTA的路由表

```
[RTA]display ip routing-table
Routing Tables: Public
        Destinations : 11       Routes : 11

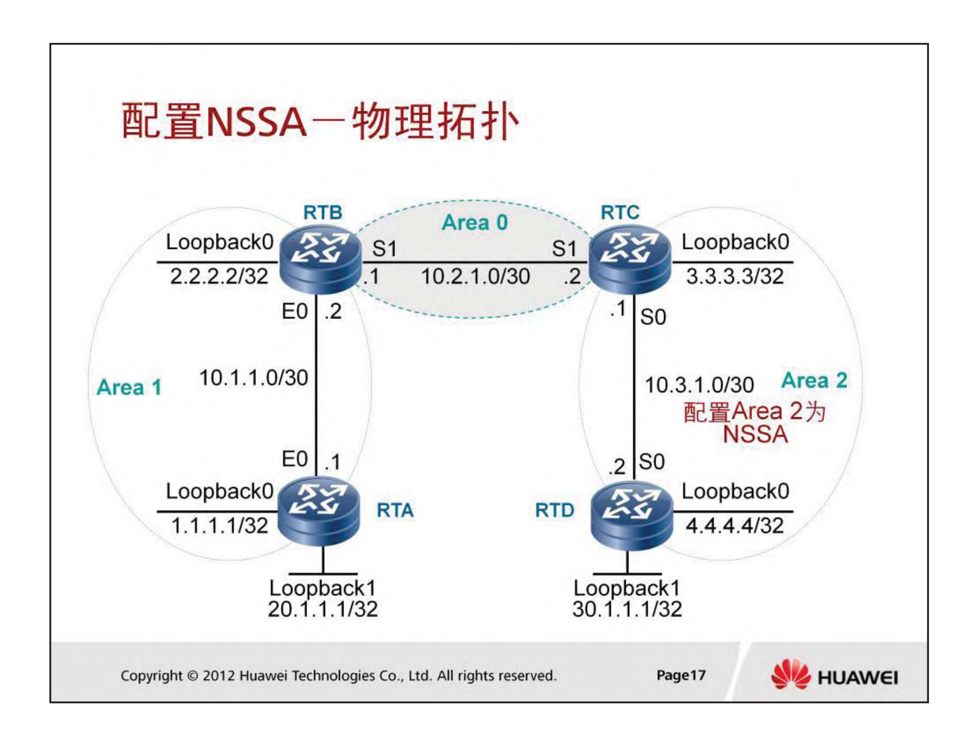
Destination/Mask    Proto  Pre  Cost       NextHop         Interface

       0.0.0.0/0    OSPF   10   2          10.1.1.2        Ethernet0
       1.1.1.1/32   Direct 0    0          127.0.0.1       InLoopBack0
      10.1.1.0/30   Direct 0    0          10.1.1.1        Ethernet0
      10.1.1.1/32   Direct 0    0          127.0.0.1       InLoopBack0
      10.5.1.0/30   Direct 0    0          10.5.1.1        Serial1
      10.5.1.1/32   Direct 0    0          127.0.0.1       InLoopBack0
      10.5.1.2/32   Direct 0    0          10.5.1.2        Serial1
      20.1.1.1/32   Direct 0    0          127.0.0.1       InLoopBack0
      20.1.1.2/32   Direct 0    0          127.0.0.1       InLoopBack0
     127.0.0.0/8    Direct 0    0          127.0.0.1       InLoopBack0
     127.0.0.1/32   Direct 0    0          127.0.0.1       InLoopBack0
```

只有默认路由，没有明细AS外部路由和明细区域间路由

 Page16 HUAWEI

在RTA（完全Stub区域的区域内路由器）的路由表中，通过OSPF学习到的只有一条ABR通告的默认路由和区域内部路由，没有区域间路由和AS外部路由。



本例中:

在RTA和RTD上引入直连路由，所以RTA和RTD是ASBR。

配置Area 2为NSSA，RTC将通告默认路由到该区域中。在RTD上引入的外部路由可以泛洪到整个AS，但是在RTA上引入的外部路由不能泛洪到Area 2中。

配置Area 1为普通区域，所以在RTA和RTD上引入的外部路由都可以被泛洪到Area 1中。

配置NSSA－配置RTA和RTB

```
[RTA]router id 1.1.1.1
[RTA]ospf
[RTA-ospf-1]area 1
[RTA-ospf-1-area-0.0.0.1]network 1.1.1.1 0.0.0.0
[RTA-ospf-1-area-0.0.0.1]network 10.1.1.0 0.0.0.255
[RTA-ospf-1-area-0.0.0.1]quit
[RTA-ospf-1]import-route direct
[RTA-ospf-1]quit
[RTA]
```

把到达**20.1.1.1/32**的路由作为外部路由引入

```
[RTB]router id 2.2.2.2
[RTB]ospf
[RTB-ospf-1]area 1
[RTB-ospf-1-area-0.0.0.1]network 2.2.2.2 0.0.0.0
[RTB-ospf-1-area-0.0.0.1]network 10.1.1.0 0.0.0.255
[RTB-ospf-1-area-0.0.0.1]quit
[RTB-ospf-1]area 0
[RTB-ospf-1-area-0.0.0.0]network 10.2.1.0 0.0.0.255
[RTB-ospf-1-area-0.0.0.0]return
<RTB>
```

 Page18 HUAWEI

在RTA上，将网段1.1.1.1/32和10.1.1.0/24宣告到Area 1中。

通过引入直连路由的方式把网段20.1.1.1/32作为AS外部路由引入。

配置RTB为ABR。

配置NSSA—配置RTC

```
[RTC]router id 3.3.3.3
[RTC]ospf
[RTC-ospf-1]area 0
[RTC-ospf-1-area-0.0.0.0]network 10.2.1.0 0.0.0.255
[RTC-ospf-1-area-0.0.0.0]quit
[RTC-ospf-1]area 2
[RTC-ospf-1-area-0.0.0.2]network 3.3.3.3 0.0.0.0
[RTC-ospf-1-area-0.0.0.2]network 10.3.1.0 0.0.0.255
[RTC-ospf-1-area-0.0.0.2]nssa default-route-advertise no-summary
[RTC-ospf-1-area-0.0.0.2]return
<RTC>

[RTD]router id 4.4.4.4
[RTD]ospf
[RTD-ospf-1]area 2
[RTD-ospf-1-area-0.0.0.2]network 4.4.4.4 0.0.0.0
[RTD-ospf-1-area-0.0.0.2]network 10.3.1.0 0.0.0.255
[RTD-ospf-1-area-0.0.0.2]nssa
[RTD-ospf-1-area-0.0.0.2]quit
[RTD-ospf-1]import-route direct
[RTD-ospf-1]quit
[RTD]
```

 Page19 HUAWEI

RTC是NSSA（Area 2）的ABR，配置Area 2为NSSA，由于第五类LSA不能被通告到该区域内，因此需要开启向区域内通告默认路由的功能。

在RTD上，配置Area 2为NSSA，通过引入直连路由的方式把30.1.1.1/32作为外部路由引入。

nssa [default-route-advertise] [no-import-route] [no-summary] [flush-waiting-timer] [set-n-bit]

default-route-advertise：该参数只用于NSSA区域的ABR或ASBR才有意义，配置后，对于ABR，不论本地是否存在缺省路由，都将生成一条Type-7 LSA向区域内发布缺省路由；对于ASBR，只有当本地存在缺省路由时，才产生Type-7 LSA向区域内发布缺省路由。

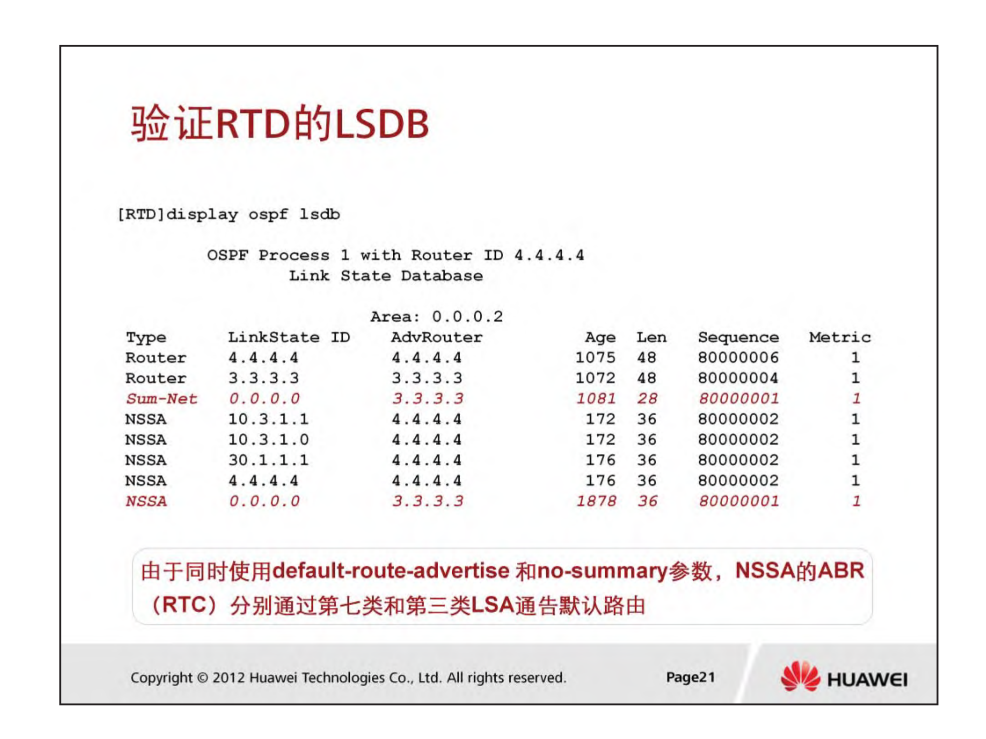
no-import-route：该参数用于禁止将AS外部路由以Type-7 LSA的形式引入到NSSA区域中，这个参数通常只用在既是NSSA区域的ABR，也是OSPF自治系统的ASBR的路由器上，以保证所有外部路由信息能正确地进入OSPF路由域。

no-summary：此参数表示不向区域内通告第三类和第四类LSA，因此NSSA中将没有区域间路由信息（类似于完全Stub区域），使用此参数之后，ABR会使用一条第三类LSA向NSSA中通告默认路由。

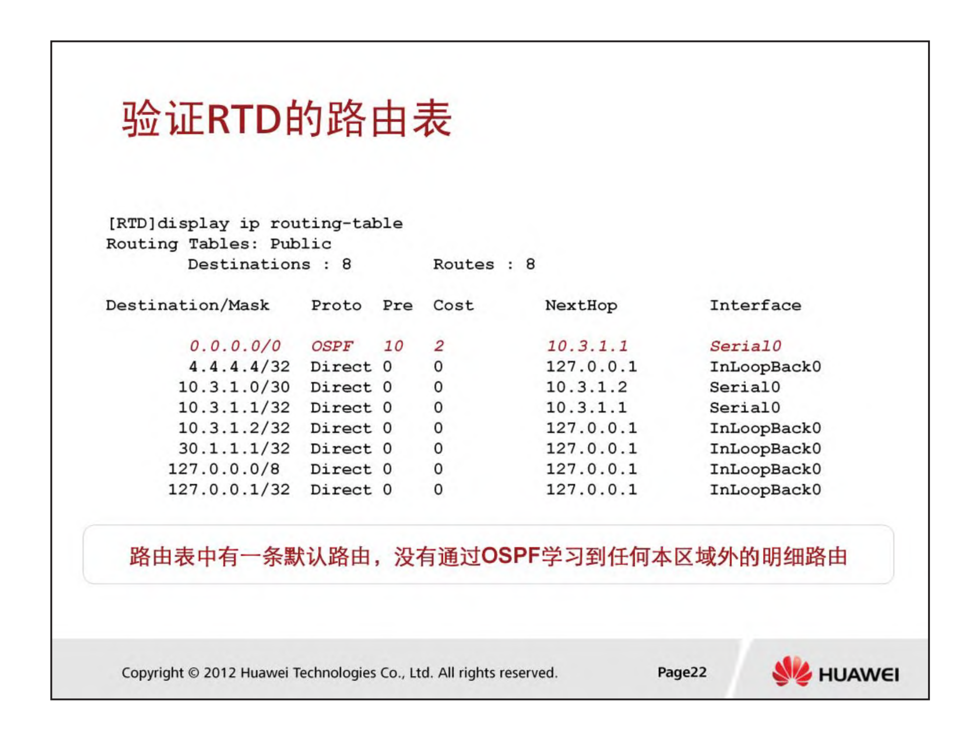
如果同时使用default-route-advertise和no-summary参数，则ABR会同时使用一条第三类LSA和一条第七类LSA分别向NSSA区域内通告默认路由，并且该默认路由只在NSSA区域内泛洪。

flush-waiting-timer：指定ASBR发送Type5 LSA的时间。

set-n-bit：在DD报文中设置N-bit位的标志。

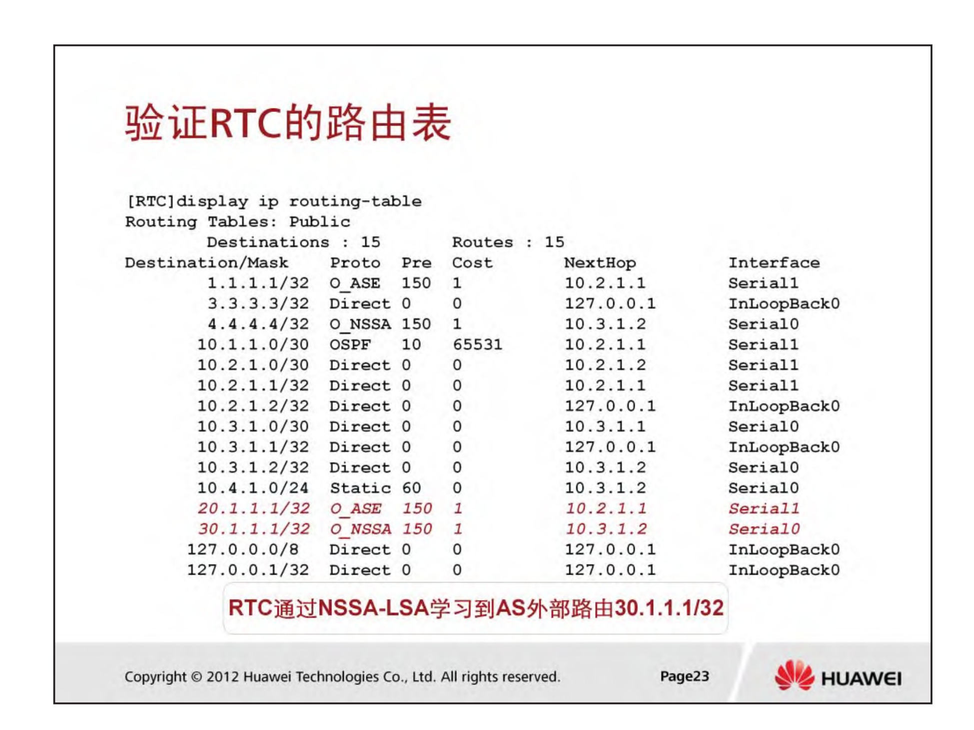


由于同时使用default-route-advertise和no-summary参数，NSSA的ABR（RTC）分别通过第七类和第三类LSA通告默认路由，通过比较优先级，使用由第三类LSA通告的默认路由（第七类LSA通告的路由被认为是AS外部路由，优先级为150）。



由于除默认路由之外，第三类、第四类和第五类LSA没有被通告到该区域中，因此RTD上没有通过OSPF学习到本区域之外的任何明细路由。

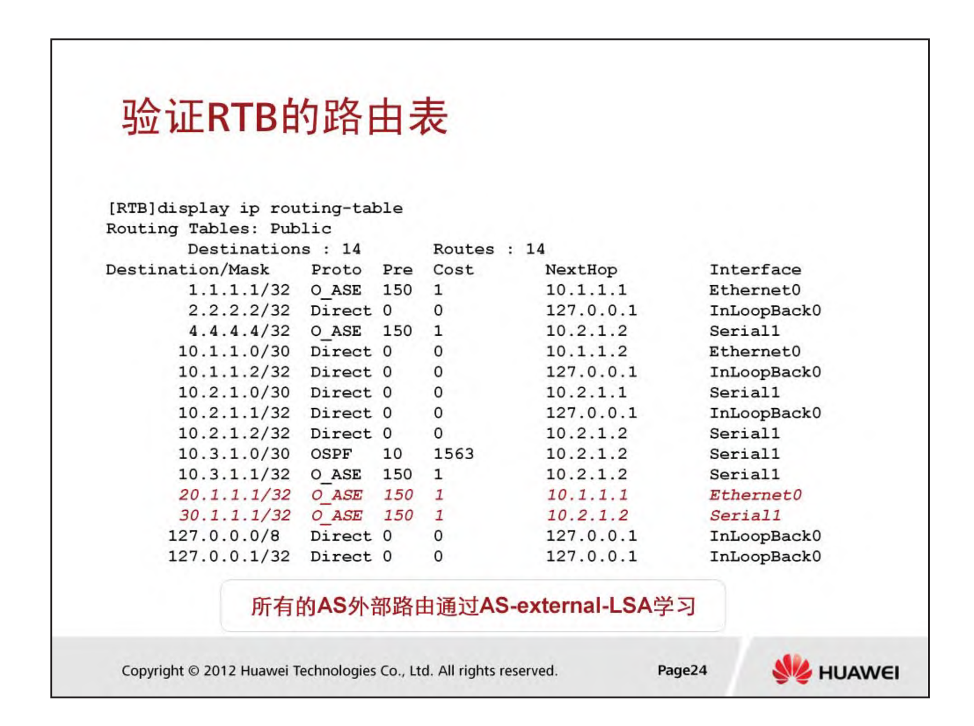
在RTA上引入的到达20.1.1.1/32的路由以及AS内的区域间路由均没有学习到。



在RTC的路由表中:

到30.1.1.1/32的路由是通过由RTD通告的一条第七类LSA学习到的。

到20.1.1.1/32的路由是通过由RTA通告的一条第五类LSA学习到的。



由于NSSA的ABR（RTC）将第七类LSA转换成了第五类LSA，所以在RTB的路由表中，所有的外部路由都是通过第五类LSA学习到的，包括在RTA上引入的到达20.1.1.1/32的路由和RTD上引入的到达30.1.1.1/32的路由。



问 题

Stub区域和完全Stub区域的区别是什么?

Stub区域和NSSA的主要区别是什么?

Page25



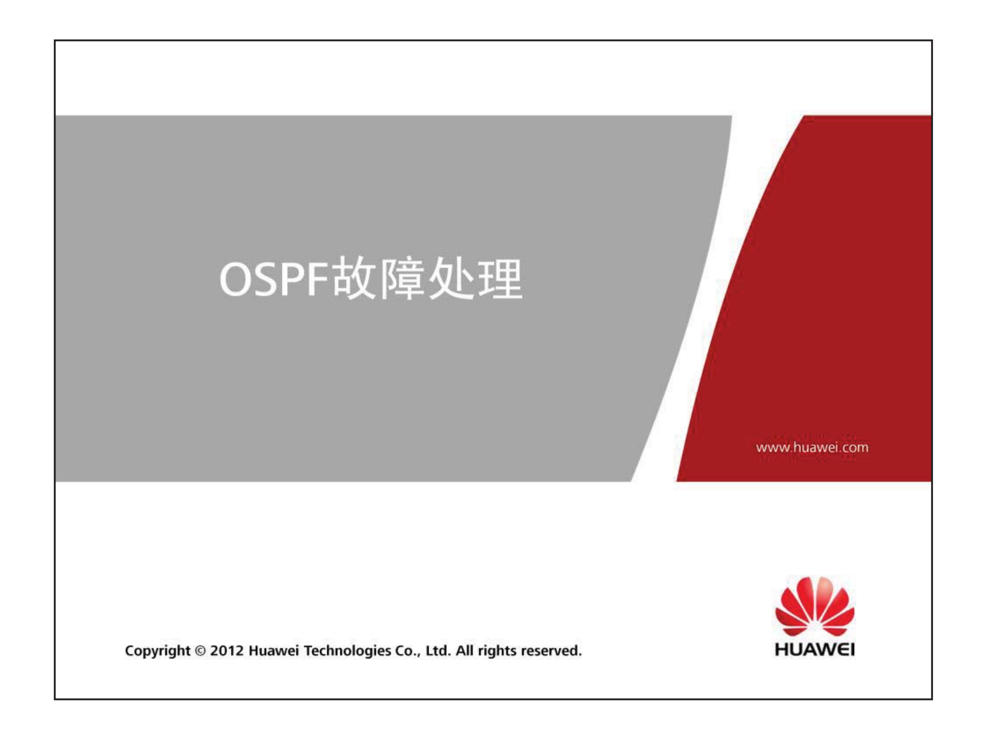
Stub区域和完全Stub区域的区别是什么?

Stub区域可以泛洪第三类和第四类LSA，完全Stub区域不可以泛洪第三类和第四类LSA（通告默认路由的第三类LSA除外）。

Stub区域和NSSA的主要区别是什么?

Stub区域不能引入AS外部路由，NSSA可以引入AS外部路由。


谢谢
www.huawei.com


OSPF故障处理
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI




前 言

本课程介绍OSPF常见故障处理。

通过处理OSPF的常见故障，可以加深对OSPF协议原理的理解。

 Page1





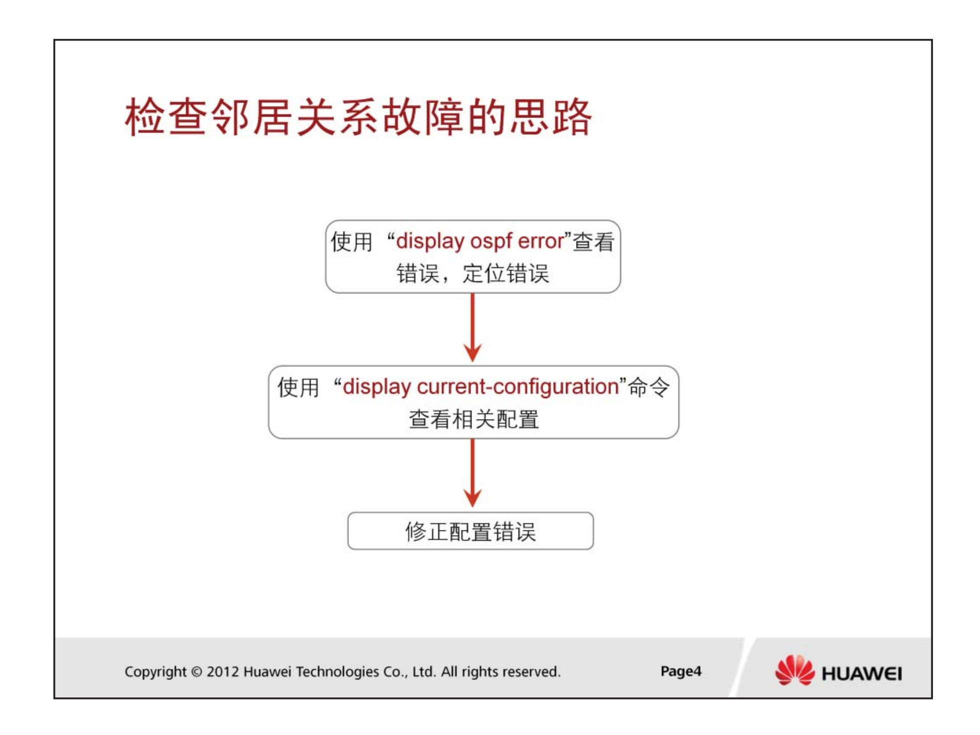
培训目标

学完本课程后，您应该能：

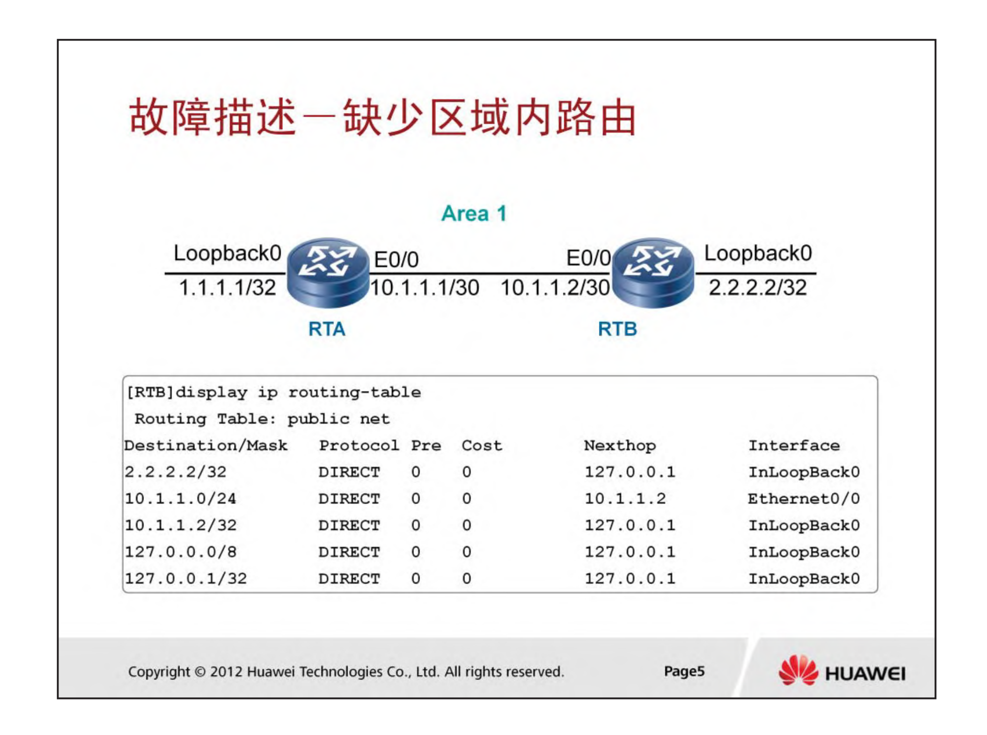
- 掌握常用OSPF故障处理工具的使用
- 掌握常见OSPF故障的处理方法




目 录
OSPF邻居关系故障
OSPF区域间路由故障
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page3
HUAWEI



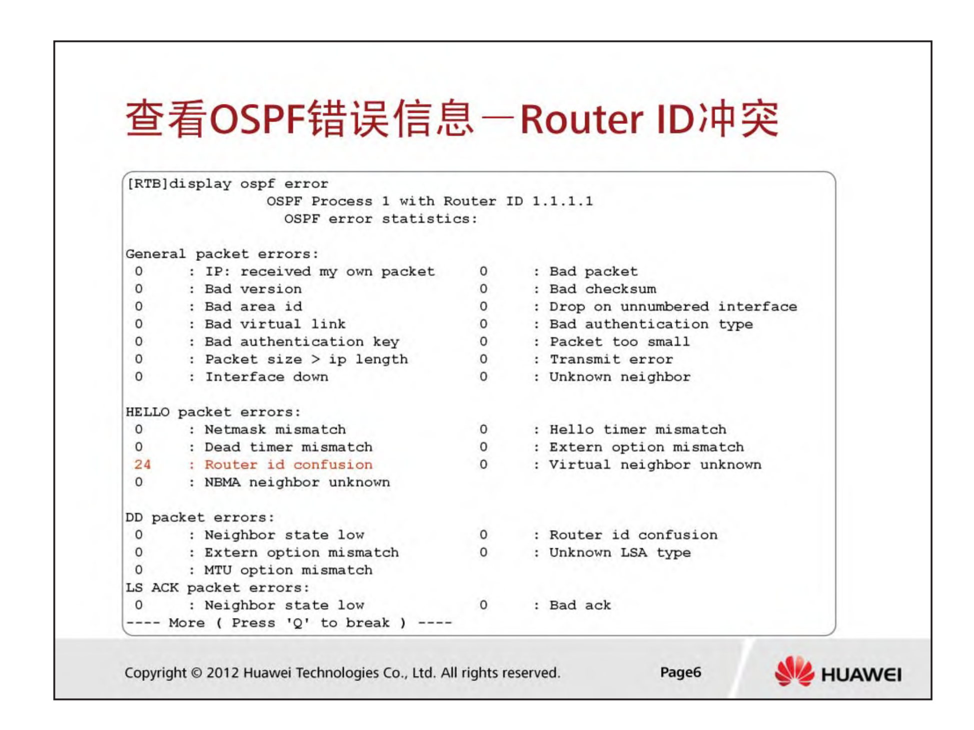
区域内路由故障通常是由于邻居关系建立错误引起的，处理这类错误的思路很简单，首先根据命令display ospf error定位错误，然后根据错误检查OSPF的相关配置，最后修正配置错误即可。



两台路由器直连，所有网段配置在同一区域中。

配置完成后，检查RTB的路由表，发现无法学习到对端Loopback接口的路由。

这种区域内缺少路由的故障，通常是由邻居关系故障引起的，处理邻居关系故障的主要思路是：首先根据命令display ospf error定位错误，然后根据错误检查OSPF的相关配置，最后修正配置错误即可。



使用命令display ospf error查看OSPF错误信息，发现有Router ID冲突的统计。

OSPF要求网络中所有的路由器都有一个Router ID，并且Router ID不得重复。

查看OSPF相关配置

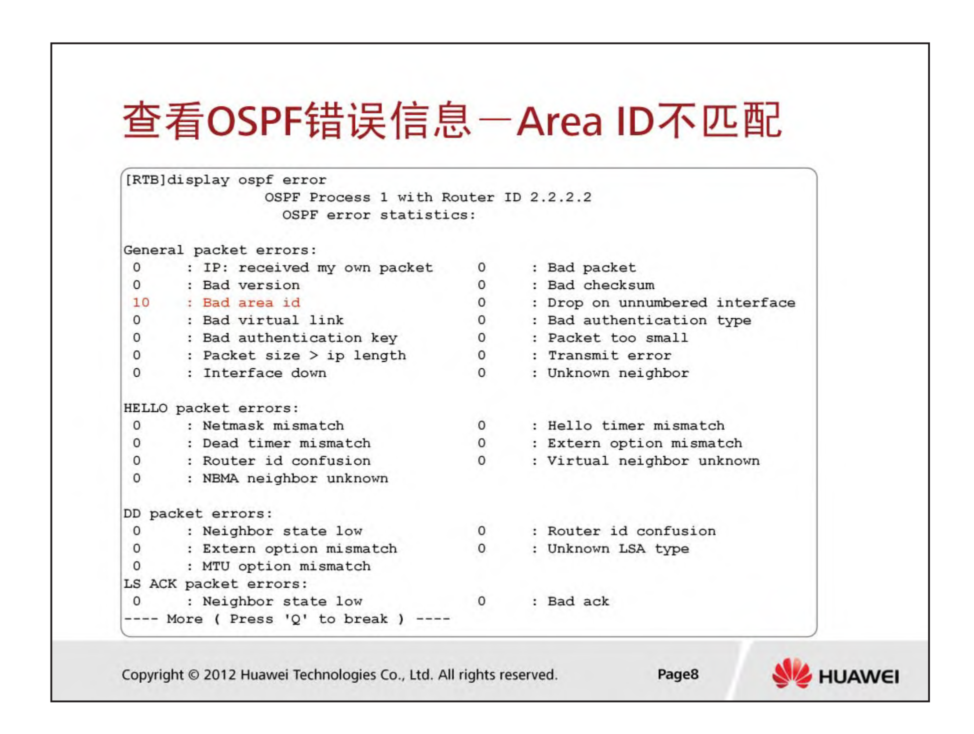
```
[RTA]display current-configuration
#
 sysname RTA
#
 FTP server enable
#
 l2tp domain suffix-separator @
#
 router id 1.1.1.1
#
radius scheme system
```

```
[RTB]display current-configuration
#
 sysname RTB
#
 FTP server enable
#
 l2tp domain suffix-separator @
#
 router id 1.1.1.1
#
radius scheme system
```

 Page7 HUAWEI

查看RTA和RTB的相关配置，发现Router ID都配置成了1.1.1.1。

修改RTB的Router ID为2.2.2.2（Loopback接口的IP地址），故障即可排除。



使用命令display ospf error查看OSPF错误信息，发现有区域id错误的统计信息。

OSPF规定，区域是一组网段的集合，即一个网段的所有接口都要配置在同一个区域中，否则无法建立邻居关系。

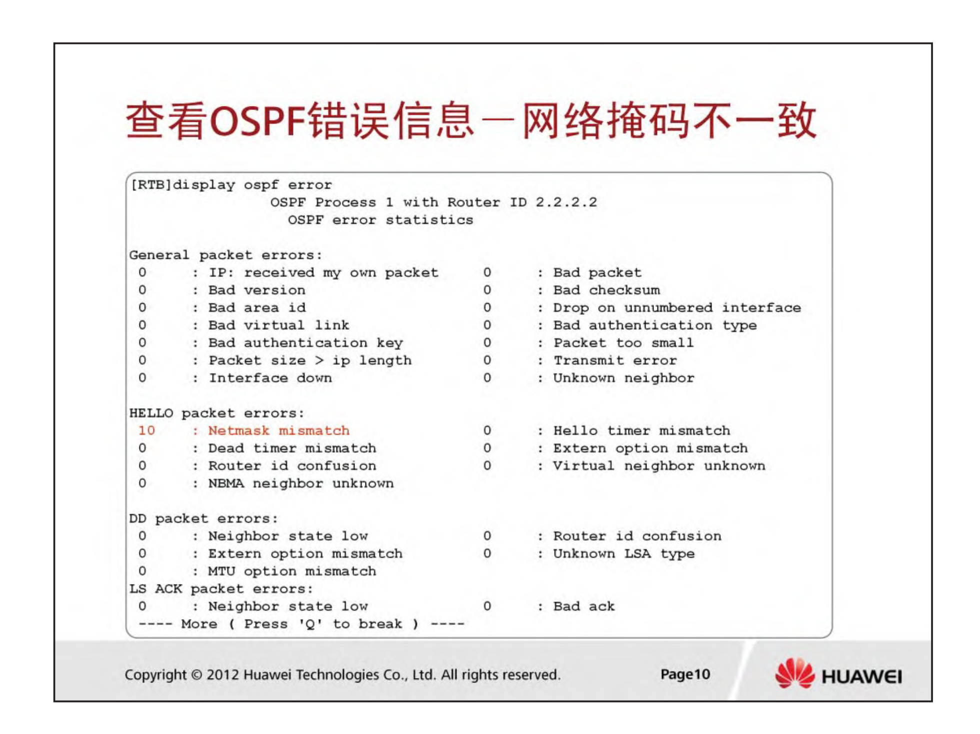
查看OSPF相关配置

```
[RTA]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.1
  network 1.1.1.1 0.0.0.0
  network 10.1.1.0 0.0.0.3
#
return
```

```
[RTB]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.2
  network 2.2.2.2 0.0.0.0
  network 10.1.1.0 0.0.0.3
#
return
```

 Page9 HUAWEI

查看OSPF相关配置，发现两个端口配置在不同的区域中。修改RTB的配置，把所有网段配置在Area 1中，故障即可排除。



使用命令display ospf error查看OSPF错误信息，发现有网络掩码不匹配的统计。

OSPF规定，在广播型、NBMA和点到多点网络中，网络掩码的配置必须相同，否则无法建立OSPF邻居关系。

查看OSPF相关配置

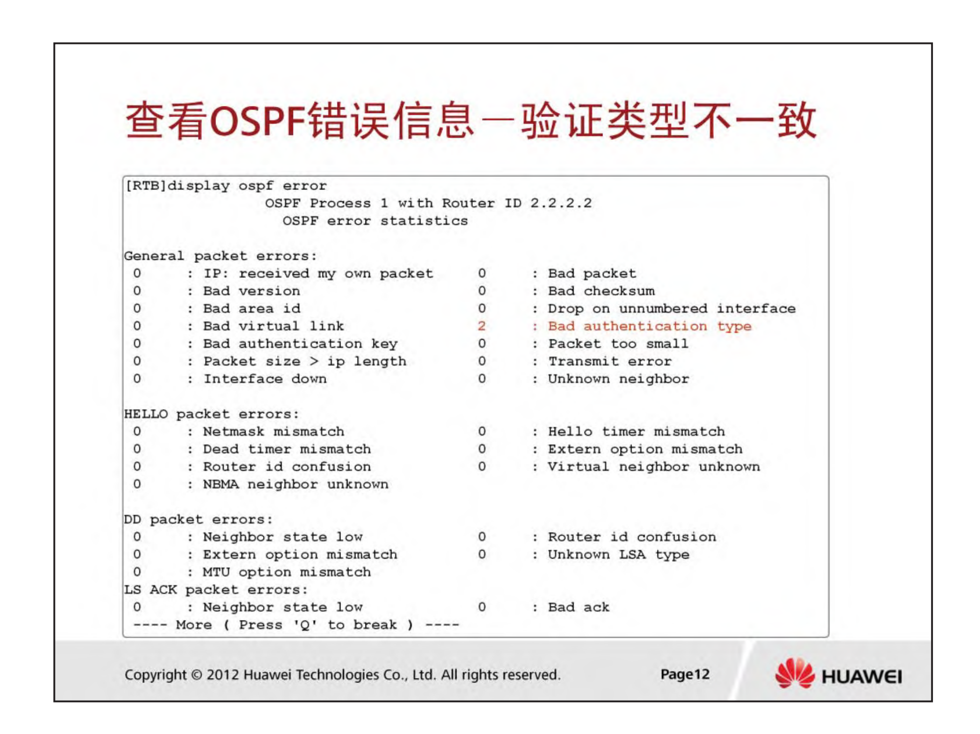
```
[RTA]display current-configuration
#
interface Ethernet0/0
 ip address 10.1.1.1 255.255.255.252
#
ospf 1
 area 0.0.0.1
  network 1.1.1.1 0.0.0.0
  network 10.1.1.0 0.0.0.3
#
return
```

```
[RTB]display current-configuration
#
interface Ethernet0/0
 ip address 10.1.1.2 255.255.255.0
#
ospf 1
 area 0.0.0.1
  network 2.2.2.2 0.0.0.0
  network 10.1.1.0 0.0.0.255
#
return
```

 Page11 HUAWEI

查看OSPF相关配置，发现RTA在E0/0接口上使用的是30位掩码，RTB使用的是24位掩码，网络掩码不匹配，邻居关系无法建立。

修改RTB的E0/0接口的网络掩码为30位；修改OSPF配置中的反码为0.0.0.3，故障即可排除。



查看OSPF错误信息，发现有认证类型错误的统计。

OSPF规定同一区域内的路由器使用同一种认证类型；认证类型在区域视图下配置。

查看OSPF相关配置

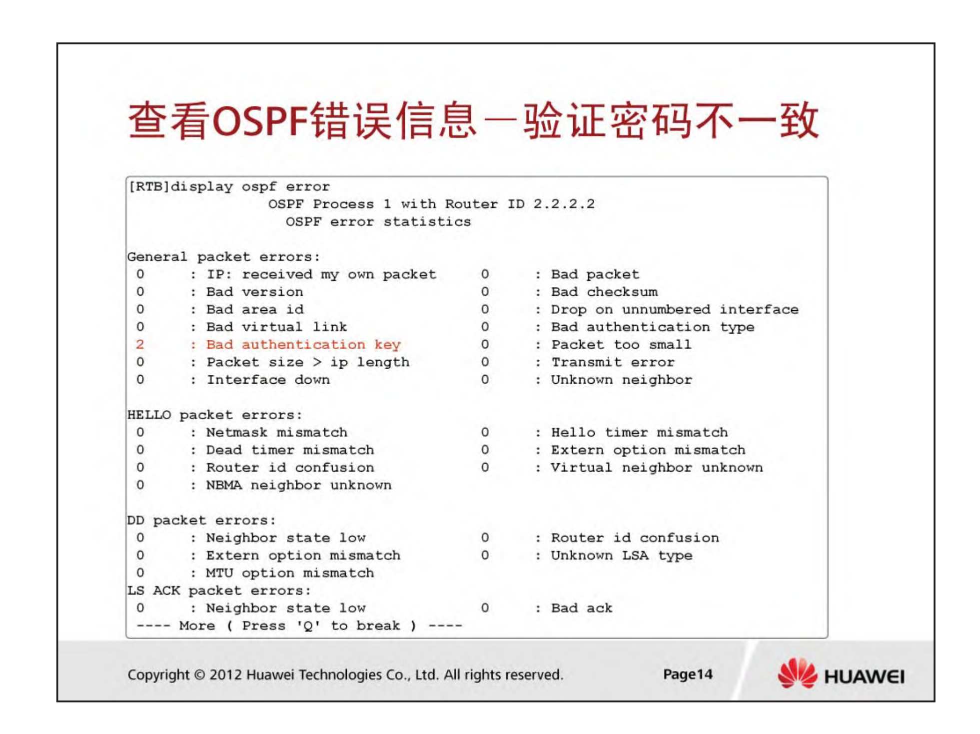
```
[RTA]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.1
  network 1.1.1.1 0.0.0.0
  network 10.1.1.0 0.0.0.3
  authentication-mode md5 1 plain huawei
#
return
```

```
[RTB]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.1
  network 2.2.2.2 0.0.0.0
  network 10.1.1.0 0.0.0.3
  authentication-mode simple plain huawei
#
return
```

 Page13 HUAWEI

查看OSPF相关配置，发现RTA使用MD5认证方式，RTB使用simple认证方式，认证方式不匹配，导致无法建立邻居关系。

修改RTA的认证类型为simple。



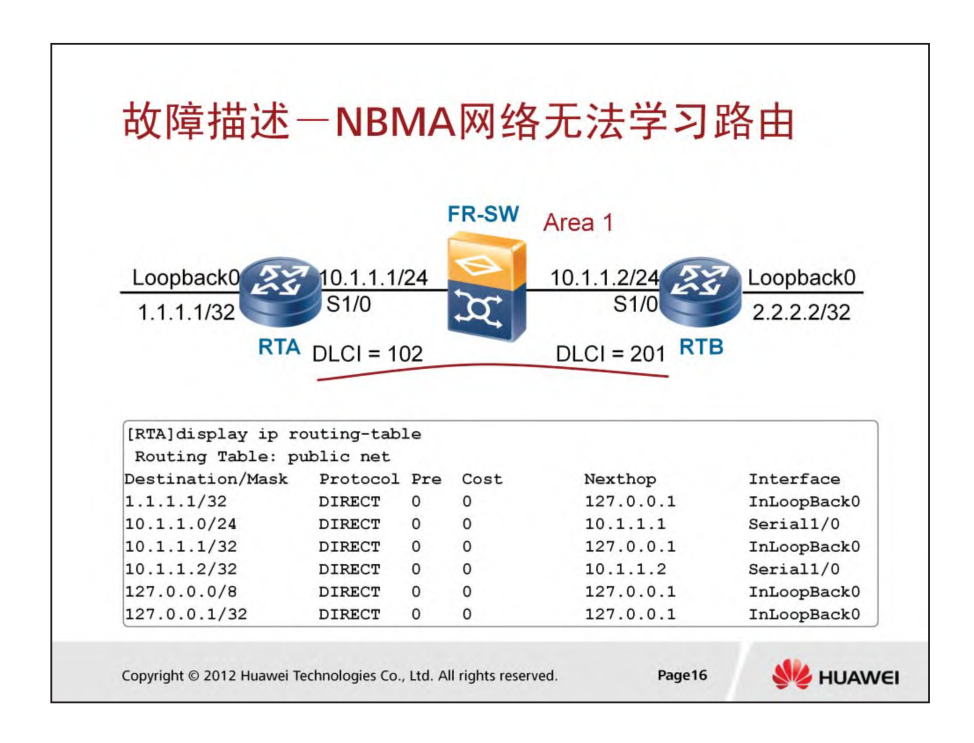
修改RTA的认证类型之后，发现仍然无法学习路由，查看OSPF错误信息，发现有验证密码不匹配的统计。

OSPF规定报文认证的密码在端口上配置，并且一条链路的两端需要配置相同的密码。



查看OSPF相关配置，发现在端口上配置的密码是不匹配的。

修改两端密码都为“huawei”，故障即可排除。



RTA和RTB通过帧中继交换机相连，网络类型配置为NBMA。

查看路由表，发现无法通过OSPF学习路由。

分析路由表，发现路由表中既有到本地端口的路由，也有到对端端口的路由，表明帧中继和IP地址配置正确，两端口可以正常通信。

NBMA网络上的邻居不能自动发现，只能通过静态配置指定邻居，此类故障首先应当查看静态邻居配置是否正确，然后如前所述，查看邻居参数是否正确。

查看NBMA静态邻居配置

```
[RTA]display current-configuration configuration ospf
#
ospf 1
 peer 10.1.1.3
 area 0.0.0.1
  network 1.1.1.1 0.0.0.0
  network 10.1.1.0 0.0.0.255
#
return
```

```
[RTB]display current-configuration configuration ospf
#
ospf 1
 peer 10.1.1.3
 area 0.0.0.1
  network 2.2.2.2 0.0.0.0
  network 10.1.1.0 0.0.0.255
#
return
```

 Page17 HUAWEI

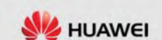
查看NBMA静态邻居的配置，发现指定对端IP地址错误。

在RTA上应使用10.1.1.2做为对端邻居的标识；在RTB上应使用10.1.1.1做为对端邻居的标识。

修改静态邻居的配置，故障即可解决。

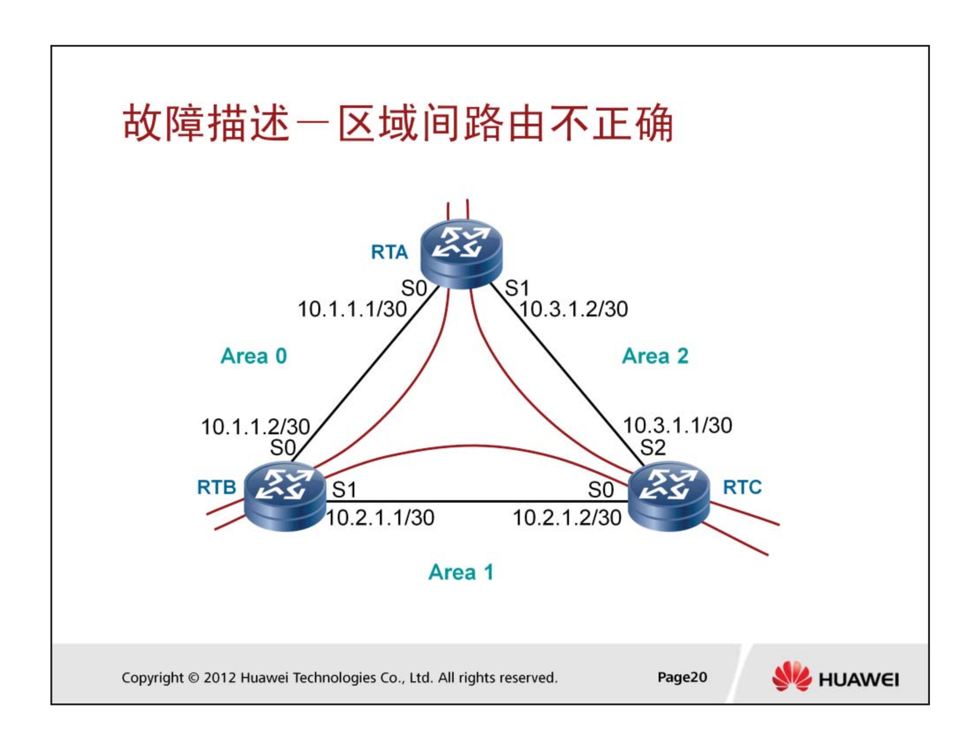
邻居关系无法建立原因总结

参数	配置要点
router id	每台OSPF路由器的router id必须唯一
area id	同一网段的所有端口应当配置在同一区域内
network mask	除了点到点网络之外，同一网段的所有端口应当配置相同的掩码
authentication type	同一区域的验证类型必须一致
authentication data	同一网段的验证码必须一致
extern option	配置stub区域或者NSSA时，区域内的所有路由器都需要指定stub特性或者NSSA特性
peer	NBMA网络上的邻居需要手动指定

 Page18 

此表格总结了引起邻居关系建立不成功的常见原因和配置要点，其中，点到点链路上不需要考虑network mask，手动指定peer只在NBMA网络上有效，其他的配置要点对所有网络类型都适用。

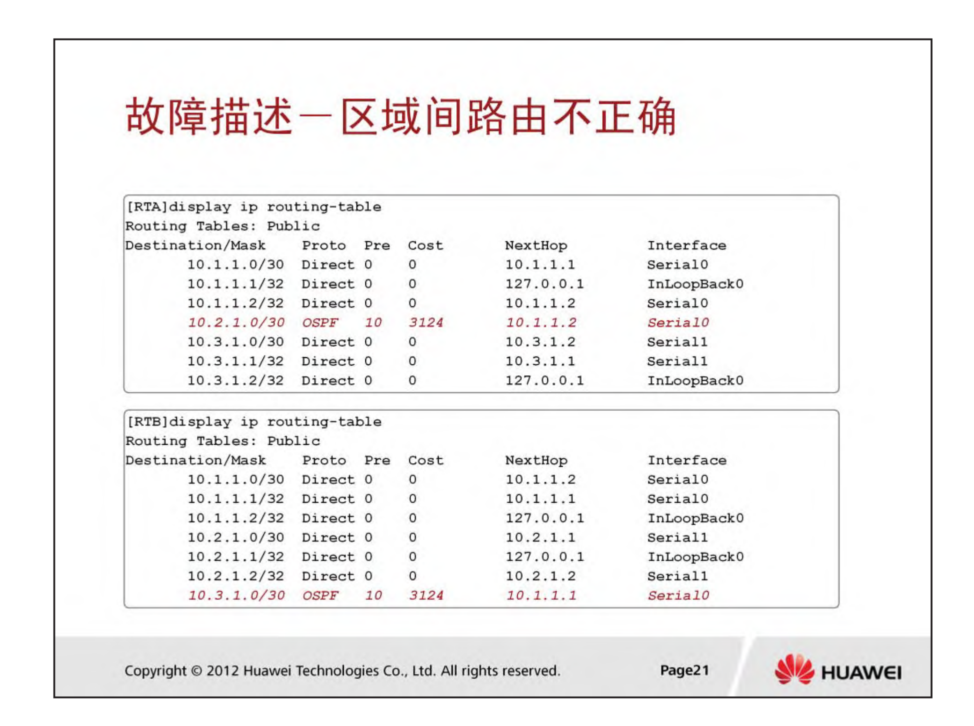

目 录
OSPF邻居关系故障
OSPF区域间路由故障
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page19
HUAWEI



本例中：

三台路由器通过串口互连，RTA使用1.1.1.1做为Router ID；RTB使用2.2.2.2做为Router ID；RTC使用3.3.3.3做为Router ID。

网段10.1.1.0/30在Area 0中，网段10.2.1.0/30在Area 1中，网段10.3.1.0/30在Area 2中，所有链路使用相同的带宽，因此每台路由器到达对端网段应当有两条等值路径。



在RTA的路由表中，到达10.2.1.0/30路径只有一条，下一跳为10.1.1.2，通过骨干区域到达10.2.1.0/30。

故障原因分析：

区域间的路由信息只能通过Area 0发布，不能在非骨干区域之间直接发布。在本例中，RTC不能将到达10.2.1.0/30的路由信息直接发布给RTA，只有Area 1的ABR（RTB）会将到达10.2.1.0/30的路由信息发布给RTA，因此在RTA的路由表中，下一跳只有10.1.1.2（RTB）。

在RTB的路由表中，到达网段10.3.1.0/30的路由中，下一跳也只有一个，下一跳为10.1.1.1（RTA）。

修改OSPF配置

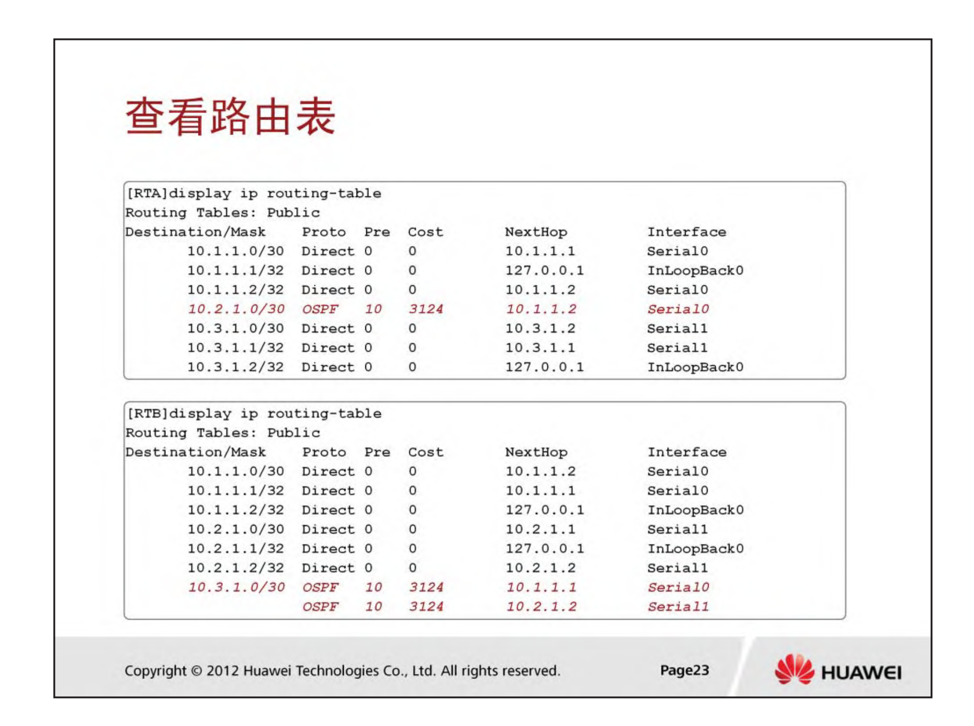
```
[RTB]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.0
  network 10.1.1.0 0.0.0.3
 area 0.0.0.1
  network 10.2.1.0 0.0.0.3
  vlink-peer 3.3.3.3
#
return
```

```
[RTC]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.1
  network 10.2.1.0 0.0.0.3
  vlink-peer 2.2.2.2
 area 0.0.0.2
  network 10.3.1.0 0.0.0.3
#
return
```

 Page22 HUAWEI

在RTB和RTC之间配置一条虚连接（Virtual Link），虚连接使用的Transit Area为Area 1。

配置虚连接的目的是为了使RTC连接到骨干区域，以使RTC可以向骨干区域发布路由信息。



修改配置之后，查看RTA和RTB的路由表，发现RTB的路由表变为正确的，在路由表中，到达10.3.1.0/30的下一跳有两个，表示两条等值路由。但是，在RTA的路由表中，到达10.2.1.0/30的下一跳仍然只有一个。

故障原因分析：

在RTB和RTC之间配置虚连接之后，Area 1是Transit区域，RTC只会把Area 2的路由信息通过Area 1发布给RTB，不会把Area 1的路由信息通过Area 2发布给RTA，因此，RTA的路由表中，到达10.2.1.0/30的下一跳仍然只有一个。

修改OSPF配置

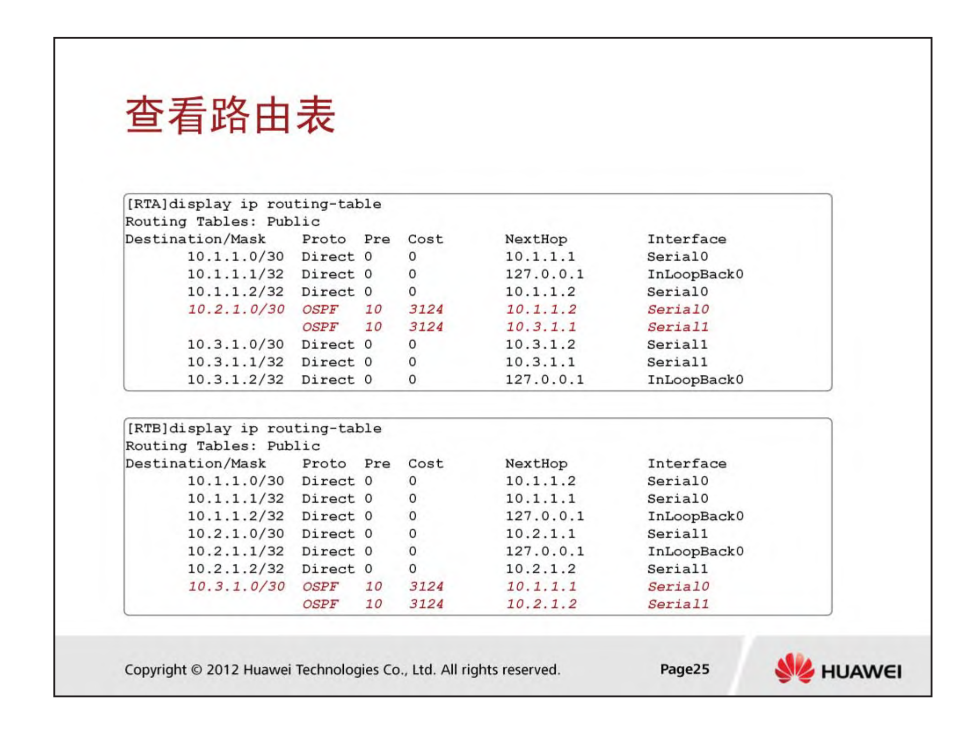
```
[RTA]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.0
  network 10.1.1.0 0.0.0.3
 area 0.0.0.2
  network 10.3.1.0 0.0.0.3
  vlink-peer 3.3.3.3
#
return
```

```
[RTC]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.1
  network 10.2.1.0 0.0.0.3
  vlink-peer 2.2.2.2
 area 0.0.0.2
  network 10.3.1.0 0.0.0.3
  vlink-peer 1.1.1.1
#
return
```

 Page24 HUAWEI

在RTA和RTC之间再配置一条虚连接（Virtual Link），虚连接使用的Transit Area为Area 2。

配置虚连接的目的是为了使RTC连接到骨干区域，同时使RTC可以向骨干区域发布Area 1和Area 2的路由信息。



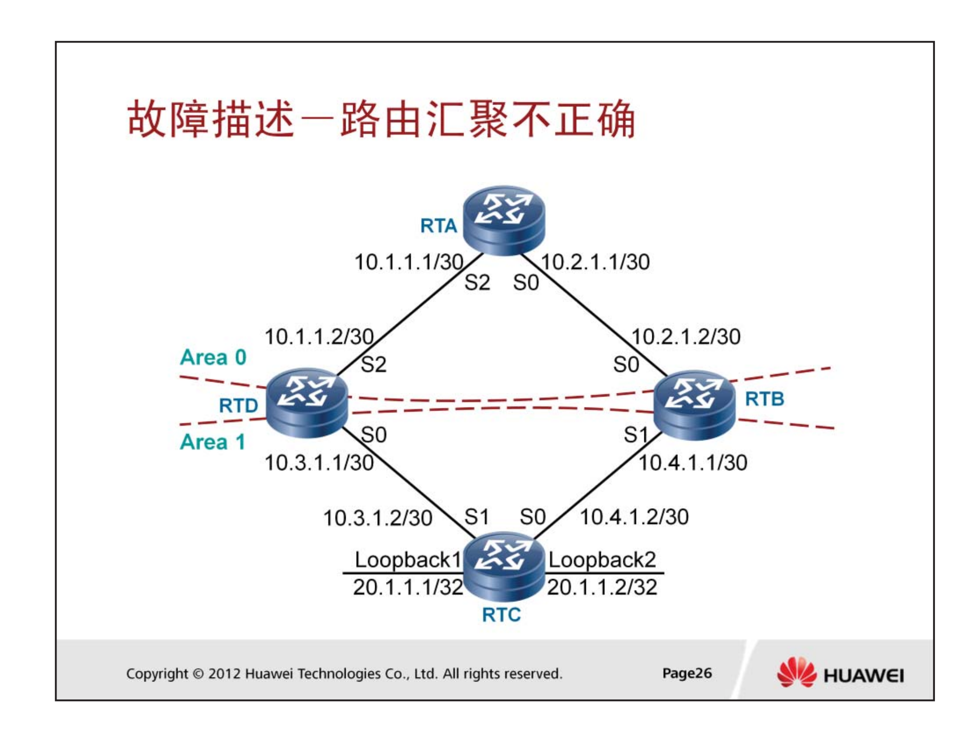
再次查看RTA和RTB的路由表，发现OSPF路由条目变为正确的，实现了等值路由。

小结：

在配置OSPF多区域时，如果一个路由器连接到两个或两个以上区域，则其中一个区域必须是骨干区域（Area 0），可以使用物理连接，也可以使用虚连接；

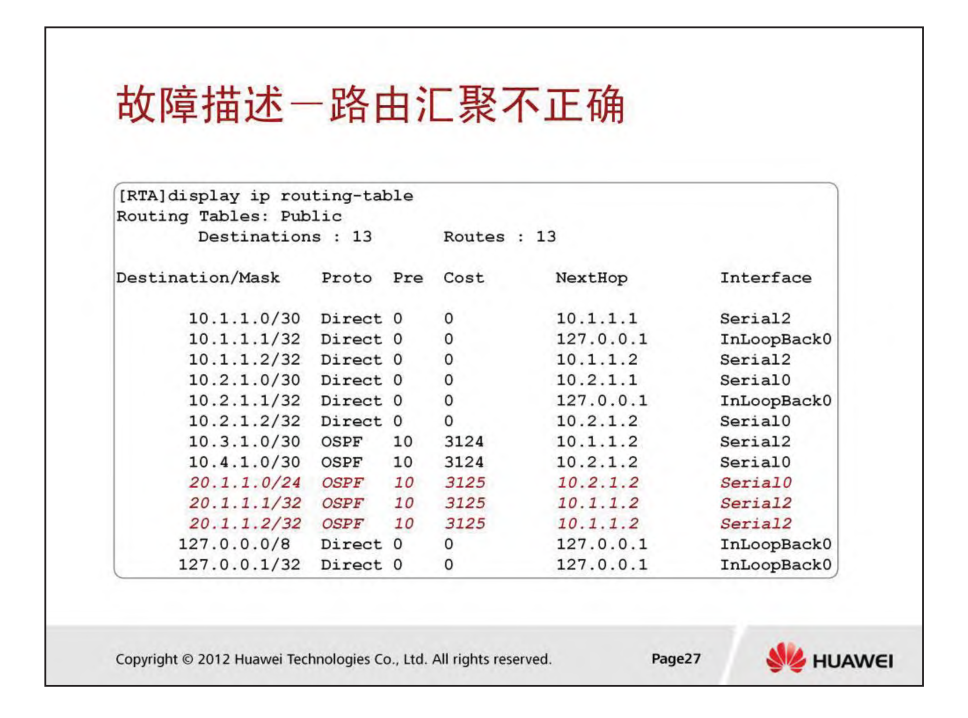
必须保证ABR把与自己相连的每一个非骨干区域的路由都发布到骨干区域中；

组网时，尽量避免使用虚连接，尽量使用以骨干区域为中心，以非骨干区域为末梢的星型结构，尽量避免非骨干区域直接相连。



本例中，在RTC上将所有网段配置在Area 1中，在Area 1的ABR上配置路由汇聚，把网段20.1.1.1/32和20.1.1.2/32汇聚成20.1.1.0/24，使ABR在向骨干区域通告时，只通告20.1.1.0/24，抑制明细路由20.1.1.1/32和20.1.1.2/32。

RTA使用1.1.1.1做为Router ID；RTB使用2.2.2.2；RTC使用3.3.3.3；RTD使用4.4.4.4。

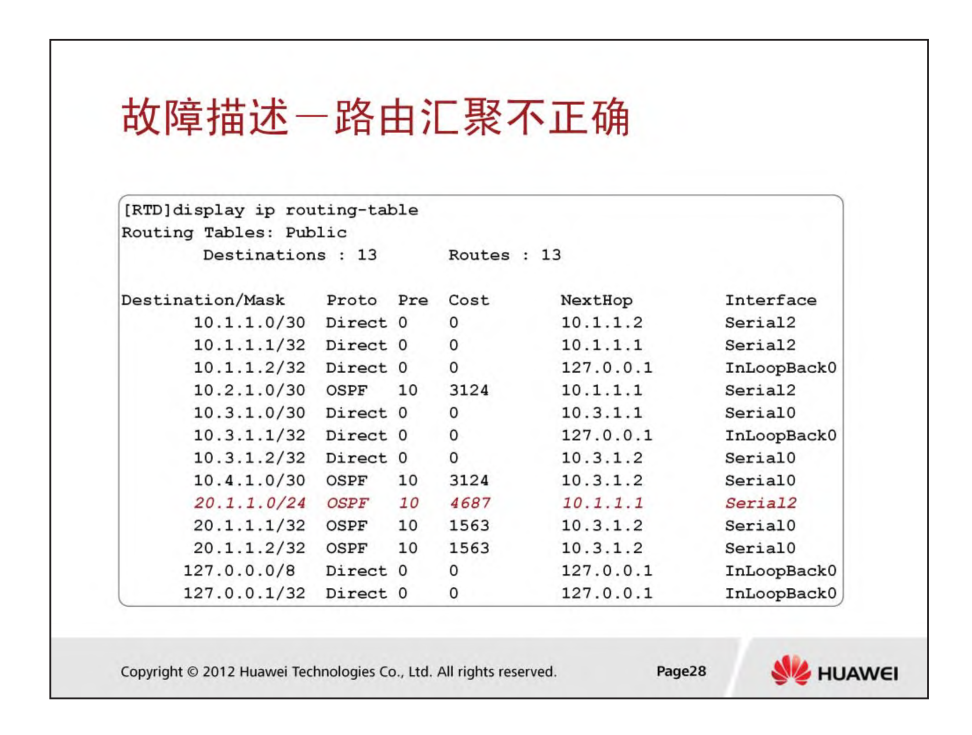


查看RTA的路由表，发现路由不正确，去往明细网段的路由使用RTD做为下一跳，去往汇聚网段的路由使用RTB做为下一跳，两者同时存在。

故障原因分析：

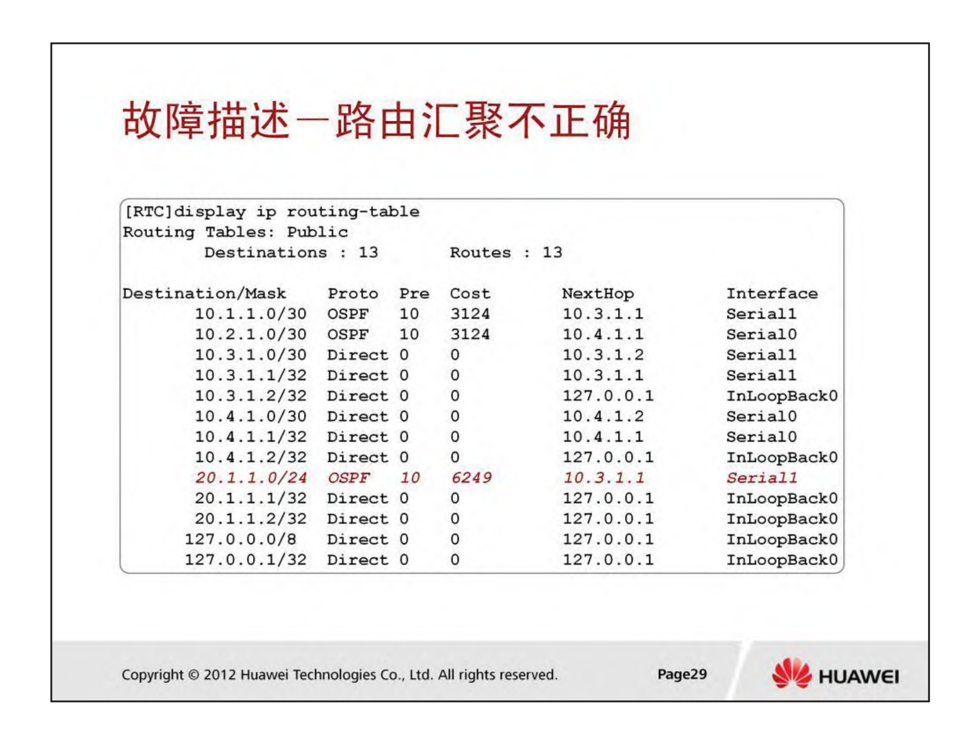
由于去往明细网段的路由使用RTD做为下一跳，说明RTD只通告了明细路由，没有通告汇聚路由，甚至可能没有做路由汇聚；汇聚路由使用RTB做为下一跳，说明RTB正确地通告了路由。

在有多个ABR地区域中配置路由汇聚地时候，所有的ABR上都要配置汇聚，不能只在一部分ABR上配置汇聚。

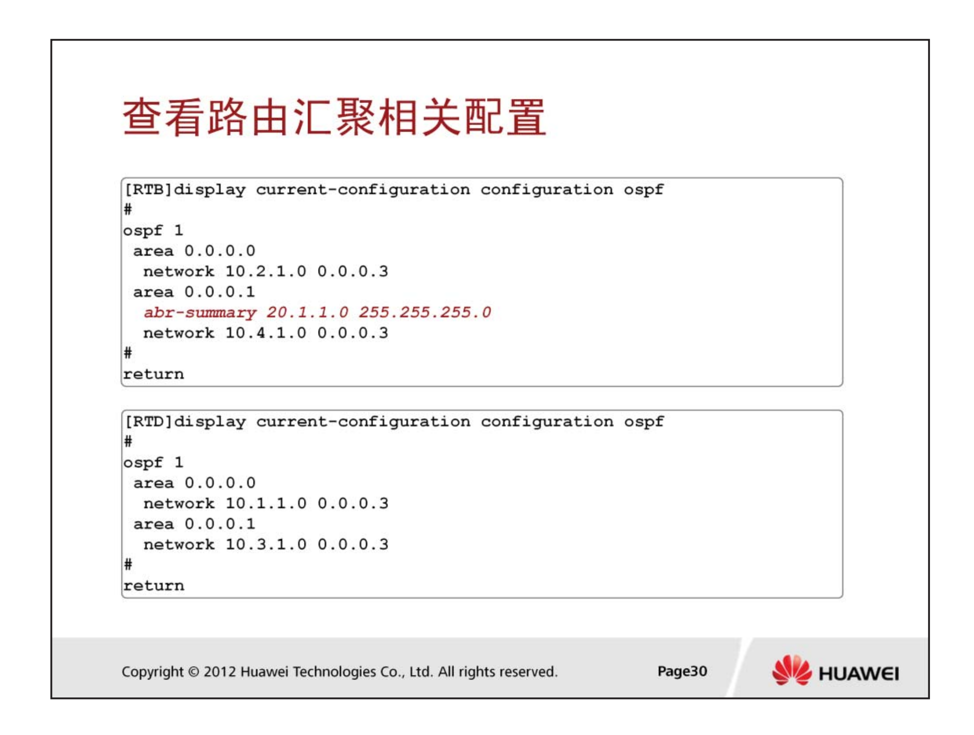


RTD的路由表中有一条汇聚路由，该路由条目是RTB通过骨干区域发布的，对于RTD来说，此路由条目是无效的。

出现此无效路由条目的原因也是RTB配置了路由汇聚而RTD没有配置路由汇聚。



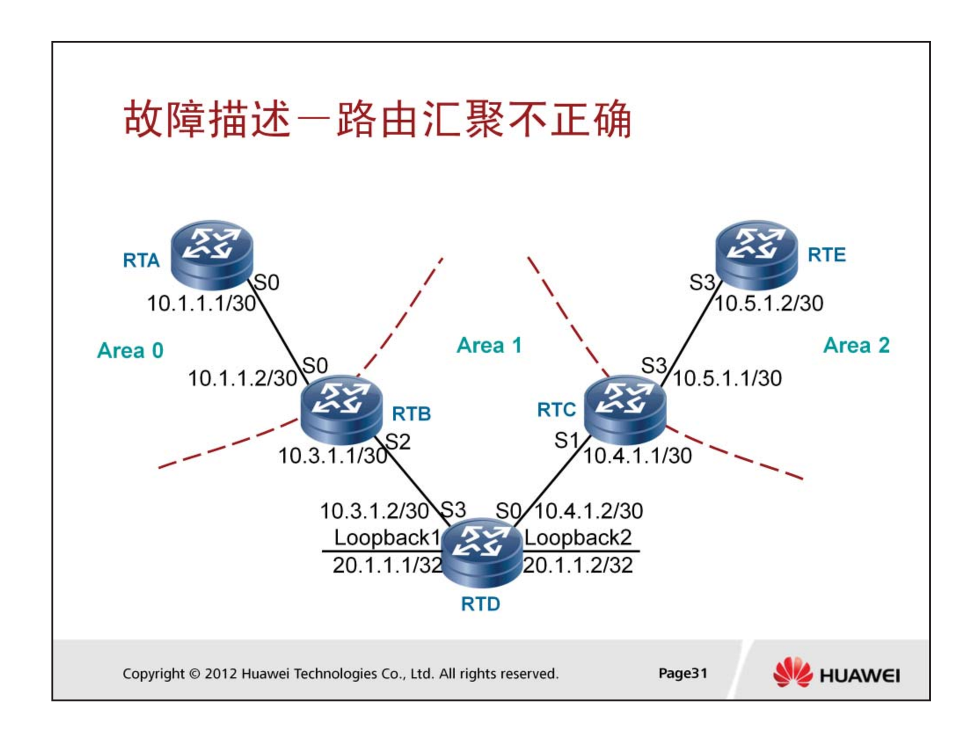
在RTC的路由表中，同样出现了一条无效路由，此路由是RTD从骨干区域接收到之后通告给RTC的。



查看两台ABR上关于OSPF的配置，发现在RTB上正确配置了路由汇聚，但是在RTD上没有配置路由汇聚。

在RTD上正确配置路由汇聚即可解决故障。

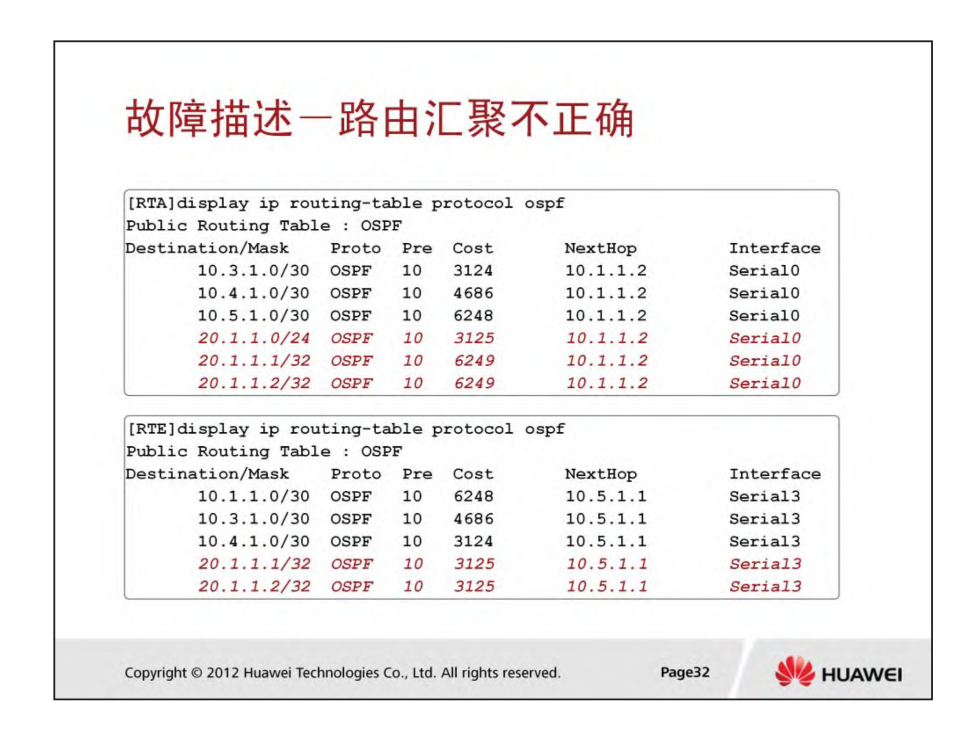
由此可见，如果路由汇聚配置不正确，足以造成路由环路。



本例中：

在RTB和RTC上配置虚连接，使Area 2连接到骨干区域。在Area 1中配置路由汇聚，把RTD上的网段20.1.1.1/32和网段20.1.1.2/32汇聚成20.1.1.0/24。

RTA使用1.1.1.1做为Router ID；RTB使用2.2.2.2；RTC使用3.3.3.3；RTD使用4.4.4.4；RTE使用5.5.5.5。



在RTA的路由表中，同时存在汇聚之后的路由和明细路由；但是在RTE的路由表中，只有明细路由没有汇聚之后的路由。

除此之外，其它的OSPF路由均正常。

故障原因分析：

由于在RTA上同时存在明细路由和汇聚之后的路由，因此考虑有可能Area 1的ABR（RTB和RTC上）上没有同时配置路由汇聚。

查看OSPF相关配置

```
[RTB]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.0
  network 10.1.1.0 0.0.0.3
 area 0.0.0.1
  abr-summary 20.1.1.0 255.255.255.0
  network 10.3.1.0 0.0.0.3
  vlink-peer 3.3.3.3
#
return
```

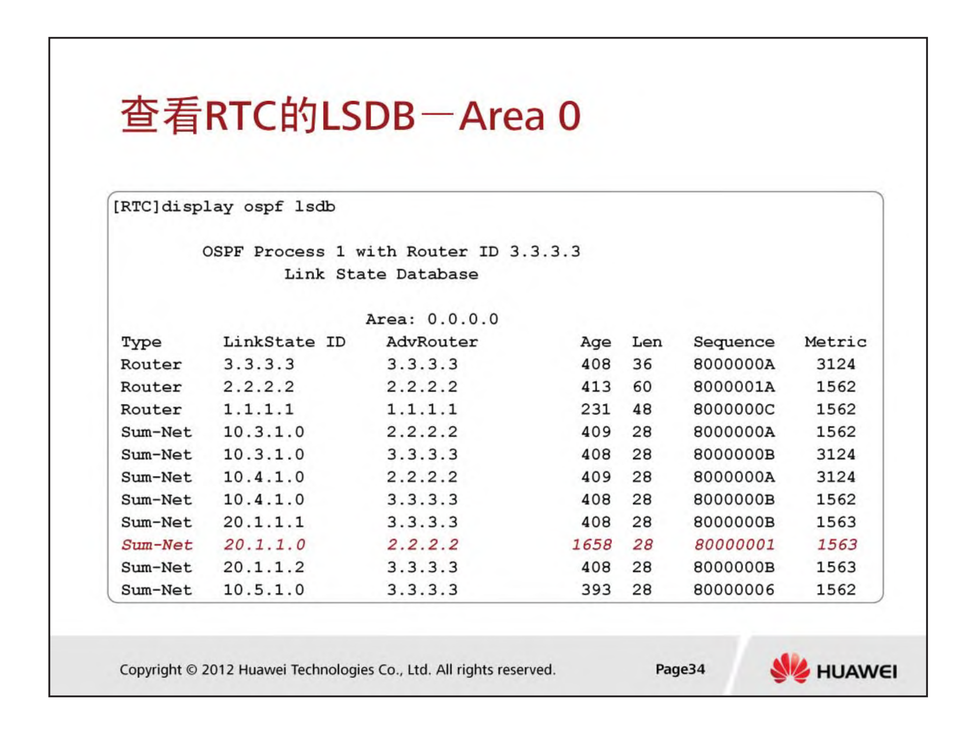
```
[RTC]display current-configuration configuration ospf
#
ospf 1
 area 0.0.0.1
  network 10.4.1.0 0.0.0.3
  vlink-peer 2.2.2.2
 area 0.0.0.2
  network 10.5.1.0 0.0.0.3
#
return
```

 Page33 HUAWEI

查看RTB和RTC的OSPF配置，发现RTB配置了路由汇聚，RTC没有配置路由汇聚，这是RTA上既有明细路由也有汇聚路由的原因，因为RTB会通过骨干区域发布汇聚路由给RTA，RTC则通过虚连接向骨干区域发布明细路由。

但是，为什么这样配置使RTE上没有汇聚路由，只有明细路由呢?

这是因为，在配置了虚连接的路由器上，如果从虚连接学习到了一条第三类LSA，只有这条LSA也从该虚连接所在的Transit区域中学习到的时候，此LSA才会被处理，也即此LSA必须在骨干区域的LSDB以及该虚连接所在的Transit区域的LSDB中同时存在，此LSA才会被处理。



RTC从骨干区域（虚连接）学到了通告汇聚之后的路由的第三类LSA，此LSA是由RTB发布的。

查看RTC的LSDB－Area 1和Area 2

```
[RTC]display ospf lsdb
        OSPF Process 1 with Router ID 3.3.3.3
                Link State Database

                        Area: 0.0.0.1
 Type      LinkState ID    AdvRouter          Age  Len   Sequence   Metric
 Router    4.4.4.4         4.4.4.4            387  96    8000001C    1562
 Router    3.3.3.3         3.3.3.3            413  48    80000018    1562
 Router    2.2.2.2         2.2.2.2            414  48    80000014    1562
 Sum-Net   10.5.1.0        3.3.3.3            395  28    80000006    1562
 Sum-Net   10.1.1.0        2.2.2.2            425  28    8000000A    1562

                        Area: 0.0.0.2
 Type      LinkState ID    AdvRouter          Age  Len   Sequence   Metric
 Router    5.5.5.5         5.5.5.5            367  48    8000000B    1562
 Router    3.3.3.3         3.3.3.3            371  48    80000007    1562
 Sum-Net   10.3.1.0        3.3.3.3            395  28    80000006    3124
 Sum-Net   10.4.1.0        3.3.3.3            395  28    80000006    1562
 Sum-Net   20.1.1.1        3.3.3.3            395  28    80000006    1563
 Sum-Net   20.1.1.2        3.3.3.3            395  28    80000006    1563
 Sum-Net   10.1.1.0        3.3.3.3            395  28    80000006    4686
```

 Page35 HUAWEI

此虚连接的Transit区域为Area 1，由于RTB汇聚的是Area 1中的路由，所以RTB不会把汇聚之后的路由发布到Area 1中，因此在Area 1的LSDB中没有通告汇聚之后的路由的第三类LSA，因此，RTC不处理这条从虚连接学习到的描述汇聚路由的第三类LSA。

由于RTC不处理从骨干区域学习到的通告汇聚路由的LSA，所以RTC没有把汇聚之后的路由发布到Area 2中，因此RTE上只有明细路由，没有汇聚路由。


问 题

导致OSPF邻居关系不能建立的原因通常有哪些?

ABR一定要连接到骨干区域吗?

路由汇聚一定要在所有ABR上配置吗?

Page36


导致OSPF邻居关系不能建立的原因通常有哪些?

Router ID冲突，Area ID不匹配，网络掩码不匹配，认证类型和认证密码不一致，外部路由能力不匹配，NBMA网络上配置了错误的静态邻居。

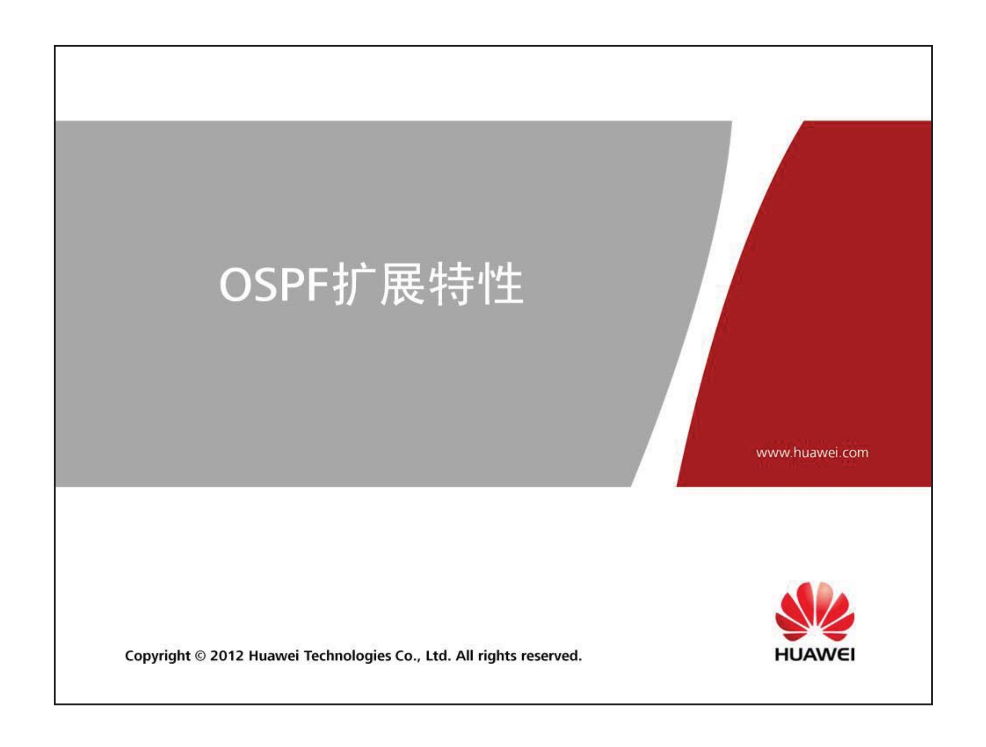
ABR一定要连接到骨干区域吗?

ABR一定要连接到骨干区域（Area 0），可以使用物理连接，也可以使用虚连接。

路由汇聚一定要在所有ABR上配置吗?

在配置路由汇聚的时候，要在被汇聚的明细网段所在的区域的所有ABR上执行。


谢谢
www.huawei.com


OSPF扩展特性
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍OSPF的一些扩展特性。

课程内容包括LSDB超载机制，按需电路（DC）特性，Stub路由器等内容。



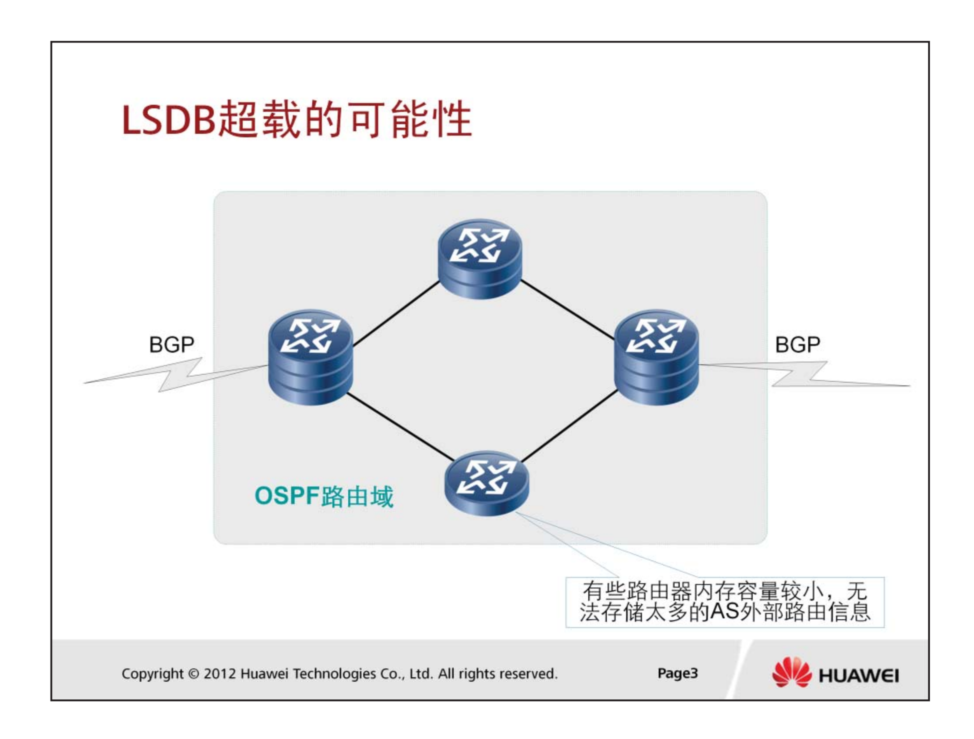


培训目标

学完本课程后，您应该能：

- 理解LSDB超载机制
- 理解按需电路上OSPF的特性
- 理解Stub路由器





由于受到存储容量的限制，当LSDB太大时，某些路由器将无法存储整个LSDB，这种现象称为LSDB超载。

LSDB超载通常是因为存储了太多AS外部路由信息（第五类LSA）引起的。

非骨干区域可以通过配置Stub区域、完全Stub区域或者NSSA减小LSDB规模。

本文介绍在不能将区域配置为Stub区域或者NSSA的情况下，使用另外一种机制处理过多的第五类LSA的方法。

LSDB超载机制

```
[RTA]ospf
[RTA-ospf-1]lsdb-overflow-limit 2
[RTA]display ospf brief

        OSPF Process 1 with Router ID 1.1.1.1
                OSPF Protocol Information

 RouterID: 1.1.1.1          Border Router:  AS
 Route Tag: 0
 Multi-VPN-Instance is not enabled
 Applications Supported: MPLS Traffic-Engineering
 Spf-schedule-interval: 5
 Default ASE parameters: Metric: 1 Tag: 1 Type: 2
 Route Preference: 10
 ASE Route Preference: 150
 SPF Computation Count: 24
 RFC 1583 Compatible
 OSPF is in LSDB overflow status
 Area Count: 1   Nssa Area Count: 0
 ExChange/Loading Neighbors: 0
```

 Page4 HUAWEI

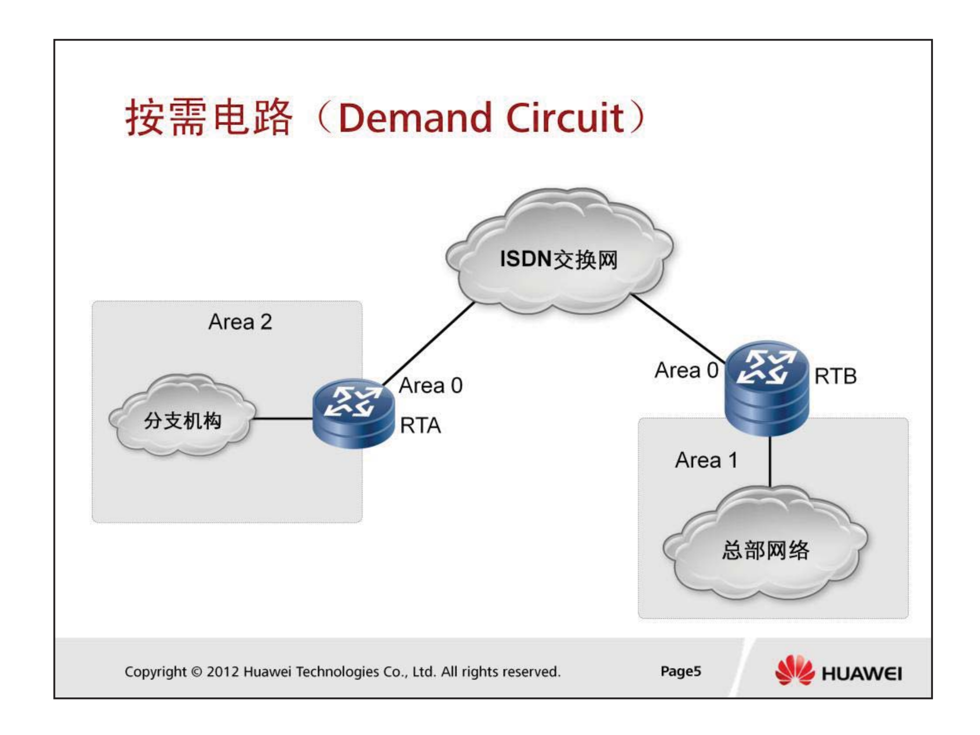
RFC1765定义了一个新的OSPF参数：ospfExtLsdbLimit，即第五类LSA的最大数量。VRP平台上的配置命令为：

lsdb-overflow-limit number

number：LSDB中第五类LSA的最大条目数，取值范围是1～1000000。默认不开启此功能。

当LSDB中的第五类LSA超过配置的最大条目数时，路由器进入超载状态，此时路由器将自身产生的第五类LSA在网络中老化，并不再生成新的第五类LSA。

处于超载状态的OSPF路由器可以接收其他路由器生成的第五类LSA，但是数量不能超过配置的最大限制，如果新接收的第五类LSA可能使LSDB中的第五类LSA超过配置的最大限制，则新接收的第五类LSA将被丢弃。



按需电路是指有流量时才会建立连接，没有流量时连接会自动断开，以便节省链路的开销。

例如，如图所示，分支机构通过ISDN交换网和总部路由器相连，使用OSPF路由协议，我们希望只在有业务数据要发送的时候，分支机构路由器才会拨号到总部网络路由器上，没有业务数据要发送的时候，拨号连接自动断开，从而节省开支。

OSPF在这种链路上不能像在普通链路上那样周期性地发送Hello报文，也不能周期性地泛洪，需要对OSPF的工作机制进行扩展。

按需电路（DC）扩展—修改Options字段

第一位	第二位	DC标志位	第四位	第五位	第六位	第七位	第八位

报文类型	DC位设置后的意义	何时设置DC位
Hello报文	用于和按需电路的对端进行协商是否在该链路上启用按需电路的扩展特性	只有按需电路上发送的Hello报文才会设置DC位
DD报文	用于和按需电路的对端进行协商是否在该链路上启用按需电路的扩展特性	只有按需电路上发送的DD报文才会设置DC位
LSA	用于通告其他路由器自己支持按需电路的扩展特性	所有自己产生的LSA都要设置此标志

Page6



第一个扩展是在OSPF的Hello报文、DD报文和LSA的Options字段中新添加一个DC标志位，标识此路由器是否支持按需电路上的扩展特性。

按需电路（DC）扩展－收发Hello报文

	普通点到点链路	按需电路（DC）
邻接关系建立以前	每10秒发送一次Hello报文	每隔Poll间隔发送一次Hello报文
邻接关系建立以后	每10秒发送一次Hello报文	不再发送Hello报文

Page7



支持按需电路扩展特性的OSPF路由器在按需电路上发送Hello报文时，要设置DC标志位；

如果支持按需电路扩展特性的OSPF路由器在某点到点链路上收到了对端发送的设置DC标志位的Hello报文，则应当认为此点到点链路为按需电路，即使在本端没有显示配置此链路为按需电路。

在普通点到点链路上，无论是否和邻居建立邻接关系，OSPF路由器都会每隔10秒（默认Hello间隔）发送一次Hello报文，但在按需电路上，OSPF对此机制做了修改：

建立邻接关系之前，每隔Poll间隔（默认120秒）发送一次Hello报文，用于检测邻居；

建立邻接关系之后，不再发送Hello报文，即始终认为对端邻居处于活动状态。

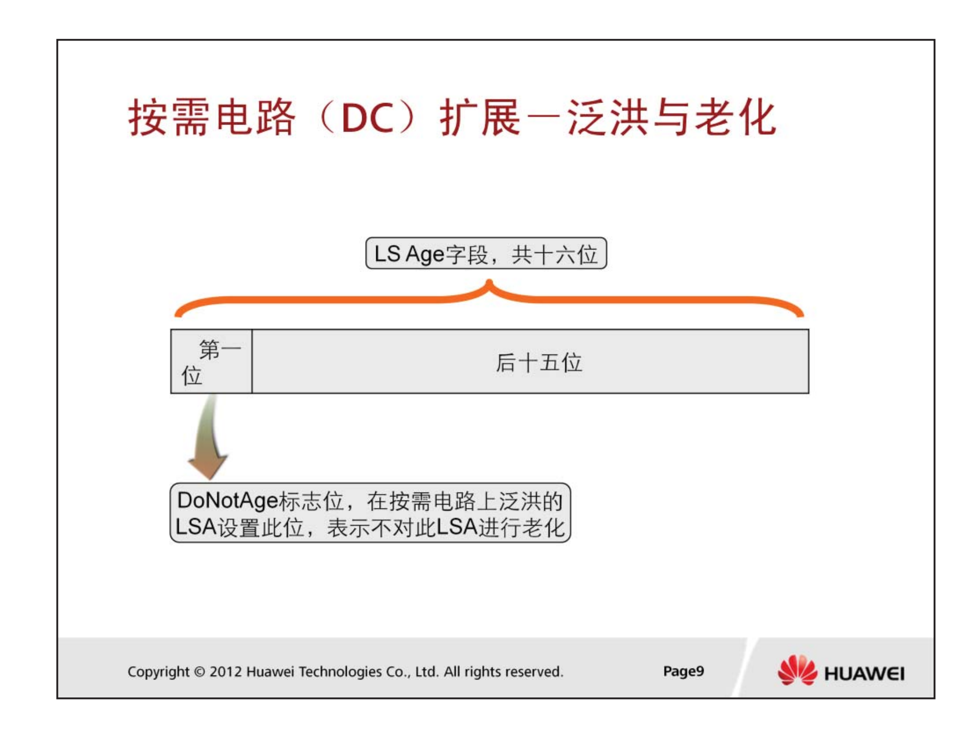
按需电路（DC）扩展－链路层始终处于Up状态

```
[RTA]display interface Dialer 0
Dialer0 current state : UP
Line protocol current state :UP (spoofing)
Description : HUAWEI, Quidway Series, Dialer0 Interface
The Maximum Transmit Unit is 1500 bytes, Hold timer is 10(sec)
Internet Address is 20.1.1.1/24
Link layer protocol is PPP
LCP initial
Physical is Dialer
    5 minutes input rate 0 bytes/sec, 0 packets/sec
    5 minutes output rate 0 bytes/sec, 0 packets/sec
    0 packets input, 0 bytes, 0 drops
    0 packets output, 0 bytes, 0 drops
```

 Page8 

按需电路接口的链路层状态始终处于Up状态，以使路由协议相信该链路处于活动状态，将此链路所处的网段向外通告。

即使物理层由于没有数据发送而自动断开，链路层仍然处于Up状态，所连接的网段仍然被路由协议向外通告。



两种情况下在按需电路上泛洪LSA：

1. LSA的内容有了真正的改变，不止是序列号或LS Age有了改变；

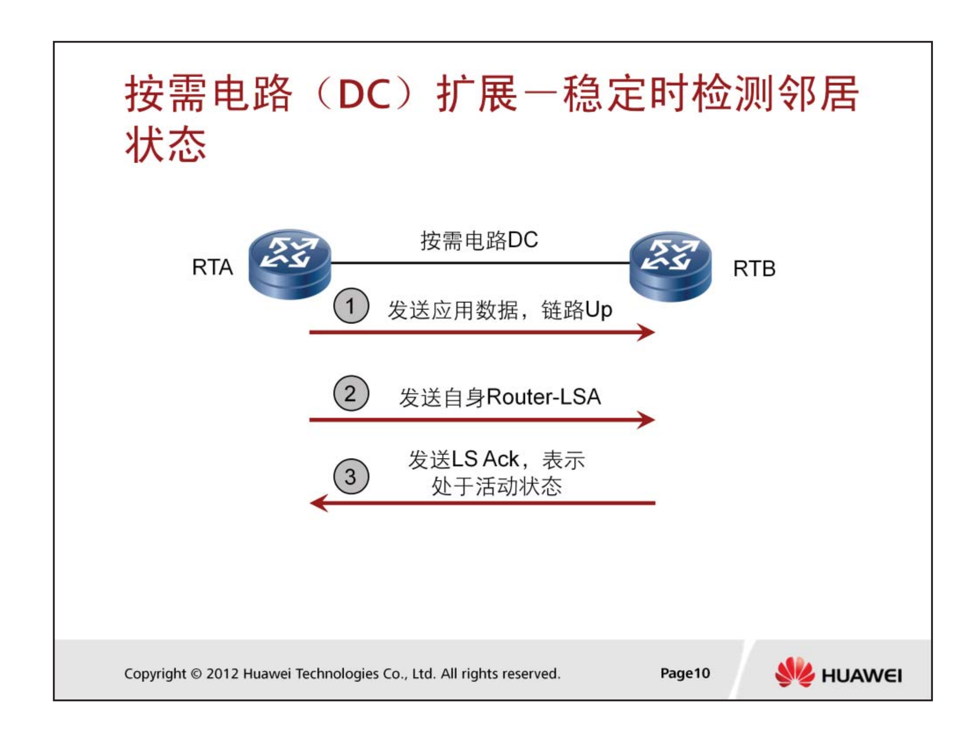
2. LSA的LS Age已经达到MaxAge（需要删除此LSA）。

由于在按需电路上发送的LSA不能像普通链路上那样可以周期性地更新，所以在按需电路扩展中规定，按需电路上发送的LSA中LS Age字段的第一位（称为DoNotAge标志位）设置为1，表示不能对此LSA进行老化。支持按需电路扩展的OSPF路由器在接收到DoNotAge标志位设置为1的LSA之后，不对此LSA进行老化。

但是在满足下列两个条件的时候，需要删除设置了DoNotAge标志位的LSA：

1. LSA在LSDB中存在了至少一小时（MaxAge）；

2. 路由表中，此LSA的生成者不可达已经至少一小时（MaxAge）。

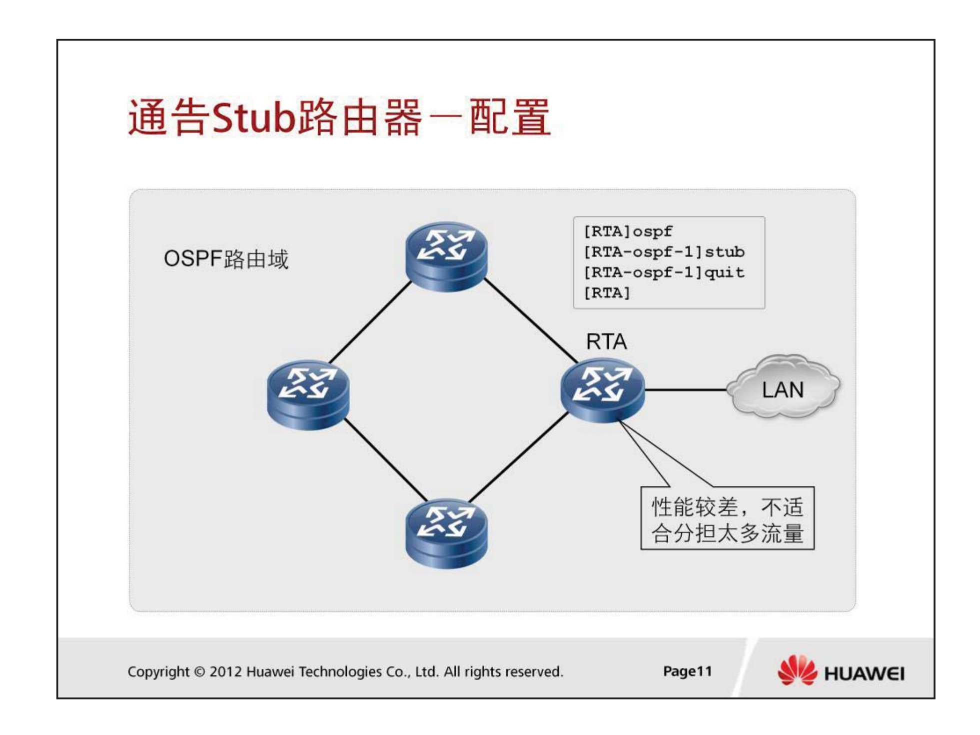


稳定状态（邻接关系已经建立的状态）下检测按需电路对端是否处于活动状态的机制如下：

1. 当有应用数据发送的时候，按需电路建立连接；

2. OSPF检测到按需电路建立连接之后，向对端发送自身生成的Router-LSA；

3. 如果对端回应LS Ack，表示对端邻居处于活动状态，如果对端在重传间隔内（默认为5秒）没有回应LS Ack，则认为对端邻居已经无效。



如图所示，RTA为低端路由器，在某些故障情况下，可能CPU占用率比较高，或者内存占用率比较高，这时管理员通常希望高速流量暂时不要经过RTA，待到故障排除之后再继续进行流量分担，同时希望到达RTA直连网段的流量不要中断。

VRP平台支持将一个路由器配置成Stub路由器，当一个路由器被配置成Stub路由器之后，在该路由器生成的Router-LSA中，非Stub连接的网段的Cost值将被通告成一个很大的值（65535），以使此链路不被优选；Stub连接的网段的Cost值不变，以使发送到直连Stub网段的数据不被中断。

stub命令用来配置此路由器为Stub路由器。

通告Stub路由器—信息显示

```
[RTA]display ospf lsdb router self-originate

        OSPF Process 1 with Router ID 1.1.1.1
                       Area: 0.0.0.0
               Link State Database

    Type       : Router
    Ls id      : 1.1.1.1
    Adv rtr    : 1.1.1.1
    Ls age     : 9
    Len        : 96
    Options    :  E
    seq#       : 8000000c
    chksum     : 0x5cf0
    Link count: 6
       Link ID: 2.2.2.2
       Data   : 10.1.1.1
       Link Type: P-2-P
       Metric : 65535
       Link ID: 4.4.4.4
       Data   : 10.4.1.1
       Link Type: P-2-P
       Metric : 65535
       Link ID: 20.1.1.0
       Data   : 255.255.255.0
       Link Type: StubNet
       Metric : 1
```

 Page12 HUAWEI

此处列出了RTA配置成Stub路由器之后，自己生成的Router-LSA的一部分信息，在自己生成的Router-LSA中，非Stub网段的Cost值被通告成了65535，直连Stub网段的Cost值没有改变。

由于非Stub网段的开销值被通告成非常大，所以其他路由器在选路的时候就会避开此路由器，于是此路由器就能尽量少的分担流量。



问 题

LSDB超载机制中限制的是第几类LSA?

按需电路上如何收发Hello报文?

按需电路上如何老化LSA?

Stub路由器连接的网段开销是多少?

 Page13



LSDB超载机制中限制的是第几类LSA?

限制的是第五类LSA。

按需电路上如何收发Hello报文?

建立邻接关系之前以Poll间隔发送Hello报文，邻接关系建立之后不再收发Hello报文。

按需电路上如何老化LSA?

设置LS Age的第一位（DoNotAge标志），以通知其他路由器不对此LSA进行老化。

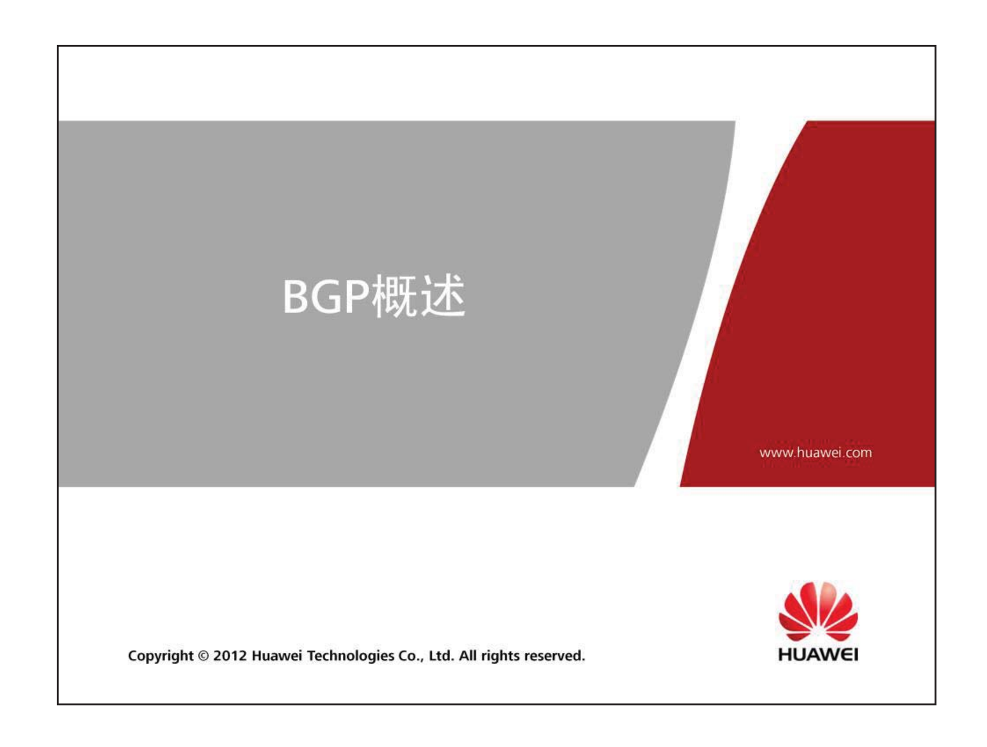
Stub路由器连接的网段开销是多少?

非Stub网段的开销值为65535，Stub网段的开销值不变。


谢谢
www.huawei.com

Module 3

BGP


BGP概述
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

动态路由协议可以按照工作范围分为IGP以及EGP。IGP工作在同一个AS内，主要用来发现和计算路由，为AS内提供路由信息的交换；而EGP工作在AS与AS之间，在AS间提供无环路的路由信息交换，BGP则是EGP的一种。





学完本课程后，您应该能：

- 知道BGP的工作范围
- 知道BGP的工作机制
- 知道BGP的特点

Page2





目 录

- 什么是BGP?
- BGP的基本工作机制
- BGP消息类型
- BGP的数据库





什么是BGP?

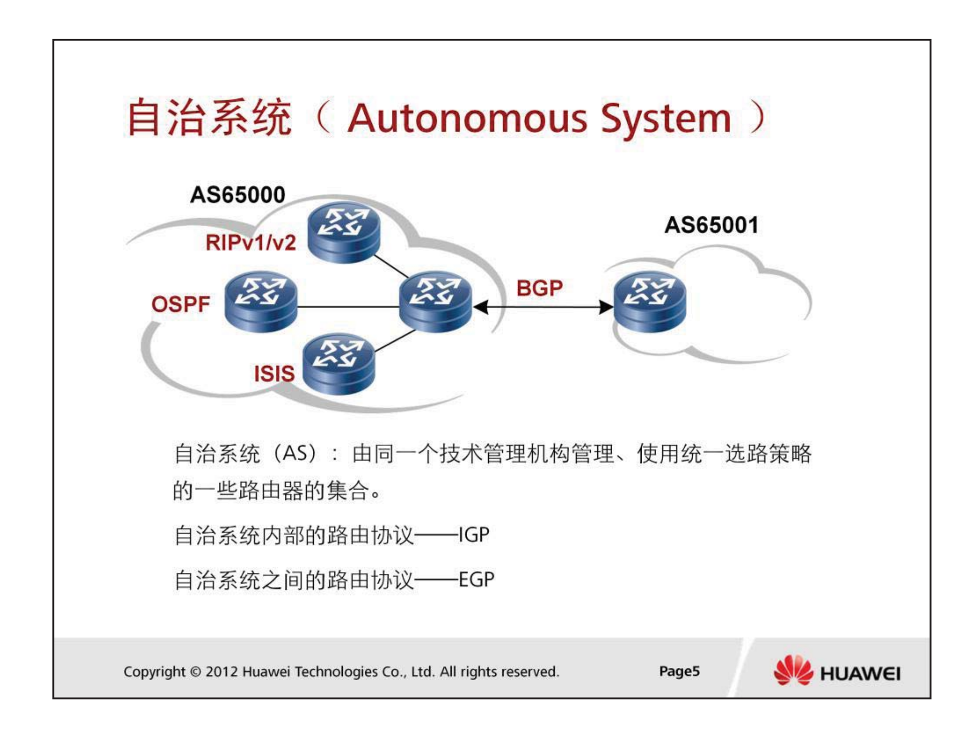
BGP的基本工作机制

BGP消息类型

BGP的数据库

Page4





自治系统的典型定义是指由同一个技术管理机构管理，使用统一选路策略的一些路由器的集合。

每个自治系统都有唯一的自治系统编号，这个编号是由IANA分配的。

我们通过不同的编号来区分不同的自治系统。当网络管理员不期望自己的数据通过某个自治系统时，比如由于该自治系统可能是由竞争对手在管理，或是缺乏足够的安全机制，因此需要回避它。这种情况下，网络管理员就可以通过路由协议、策略和自治系统编号控制数据转发的路径。

自治系统的编号范围是从1到65535，其中1到64511是注册的因特网编号，64512到65535是私有网络编号。



IGP与EGP的区别在于:

1.IGP是运行于AS内部的路由协议，主要有: RIP, OSPF及ISIS。IGP着重于发现和计算路由。

2.EGP是运行于AS之间的路由协议，现通常都是指BGP。BGP着重于控制路由的传播和选择最优的路由。

BGP 特征

BGP是外部路由协议，用来在AS之间传递路由信息

是一种增强的距离矢量路由协议

- 可靠的路由更新机制
- 丰富的Metric度量方法
- 从设计上避免了环路的发生

为路由附带属性信息

支持CIDR（无类别域间选路）

丰富的路由过滤和路由策略

 Page7 

BGP（Border Gateway Protocol）是一种自治系统间的动态路由协议，它的基本功能是在自治系统间自动交换无环路的路由信息，通过交换带有自治系统号序列属性的路径可达信息，来构造自治系统的拓扑图，从而消除路由环路并实施用户配置的路由策略。与OSPF和RIP等在自治系统内部运行的协议相比，BGP是一种EGP（Exterior Gateway Protocol），而OSPF、RIP、ISIS等为IGP（Interior Gateway Protocol）。BGP协议经常用于ISP之间。

BGP从1989年就已经开始使用。它最早发布的三个版本分别是RFC1105（BGPv1）、RFC1163（BGPv2）和RFC1267（BGPv3），当前使用的是RFC4271/RFC1771（BGPv4）。BGPv4正迅速成为Internet边界路由协议标准。

BGP的特性描述如下:

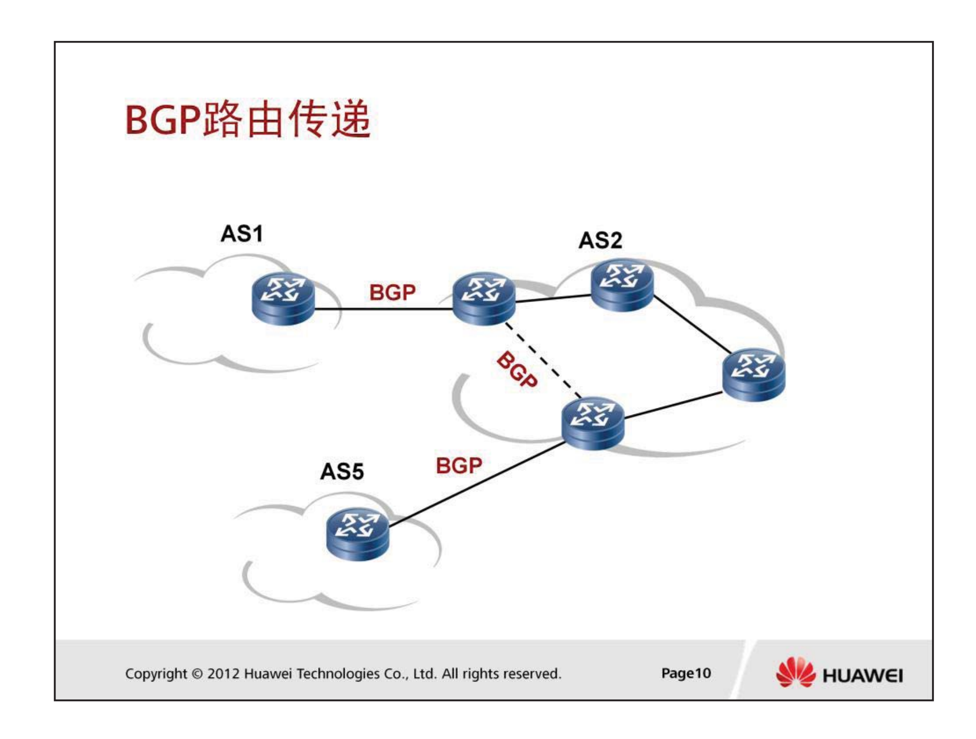
边界网关协议(BGP)，提供自治系统之间无环路的路由信息交换（无环路保证主要通过其AS-PATH实现），BGP是基于策略的路由协议，其策略通过丰富的路径属性(attributes)进行控制。BGP工作在应用层，在传输层采用可靠的TCP作为传输协议（BGP传输路由的邻居关系建立在可靠的

TCP会话的基础之上）。在路径传输方式上，BGP类似于距离矢量路由协议。而BGP路由的好坏不是基于距离（多数路由协议选路都是基于带宽的），它的选路基于丰富的路径属性，而这些属性在路由传输时携带，所以我们可以把BGP称为路径矢量路由协议。如果把自治系统浓缩成一个路由器来看待，BGP作为路径矢量路由协议这一特征便不难理解了。除此以外，BGP又具备很多链路状态（LS）路由协议的特征，比如触发式的增量更新机制，宣告路由时携带掩码等。




Page9





因为要建立TCP连接，所以两端的路由器必须知道对方的IP地址，可以通过直连端口，静态路由或者IGP学习。

ISP边界路由器知道对方的IP地址后，就可以尝试跟对方建立连接了，如果连接不能建立，说明对方还未激活，于是会等待一段时间再进行连接，这个过程一直重复，直到连接建立。

如果TCP连接建立起来，两端的设备必须交换某些数据以确认对方的能力或确定自己下一步的行动，即所谓的能力交互。这个过程是必须的，因为任何支持IP协议栈的设备都支持TCP连接的建立，但不是每个支持IP协议栈的设备都支持BGP，所以必须在该TCP连接上进行确认。

确认对方支持BGP协议后，就进行路由表的同步。两端路由表同步完成之后，并不是立即拆除这个连接。如果把这个TCP连接给拆除了，以后路由表发生改变，同步的时候就必须重新建立，这样需要消耗很多资源。如果利用保持的TCP连接，就可以不用重新建立连接而马上进行数据的传输。

建立连接的两台设备互为对等体（PEER）。为了确保两边设备的BGP进程都正在运行，要求两端的设备通过该TCP连接周期性的发送KeepAlive消息，以向对端确认自己还存活。

如果一端设备在一个存活超时的时间内没有接收到对方的KeepAlive消息，则认为对方已经停止运行BGP进程，于是拆除该TCP连接，并把从对方接收到的路由全部删除。

BGP 可靠的路由更新

传输协议：TCP，端口号179

无需周期性更新

路由更新：只发送增量路由

周期性发送keepAlive报文检测TCP的连通性

 Page12 

BGP使用TCP作为其承载协议，提高了协议的可靠性。

无需周期性发送更新消息。

路由更新时，BGP只发送增量路由（增加、修改、删除的路由信息），大大减少了传播路由时所占用的带宽，适用于在Internet上传播大量的路由信息。BGP初始化时发送所有的路由给BGP对等体，同时在本地保存已经发送给BGP对等体的路由信息。当本地的BGP收到了一条新路由时，与保存的已发送信息进行比较，如未发送过，则发送，如已发送过，则与已经发送的路由进行比较，如新路由更优，则发送此新路由，同时更新已发送信息，反之则不发送。当本地BGP发现一条路由失效时（如对应端口失效），如果路由已发送过，则向BGP对等体发送一个撤消路由的消息。总之，BGP不是每次都广播所有的路由信息，而是在初始化全部路由信息后只发送路由增量。这样保证了BGP和对端通信时占用最少的带宽。

另外，BGP通过接收和发送keepAlive消息来检测相互之间的TCP连接是否正常。



Page13


BGP报文种类

BGP报文有五种类型:

- Open：负责和对等体建立邻居关系。
- KeepAlive：该消息在对等体之间周期性地发送，用以维护连接。
- Update：该消息被用来在BGP对等体之间传递路由信息。
- Notification：当BGP Speaker检测到错误的时候，就发送该消息给对等体。
- Route-refresh：用来通知对等体自己支持路由刷新能力。

 Page14



运行BGP的路由器称之为BGP Speaker，它们之间将会交换五种类型的报文，其中OPEN报文，KEEPALIVE报文以及NOTIFICATION报文用于邻居关系的建立和维护。

Open：主要包括BGP版本，AS号等信息。试图建立BGP邻居关系的两个路由器在建立了TCP会话之后开始交换OPEN信息以确认能否形成邻居关系。

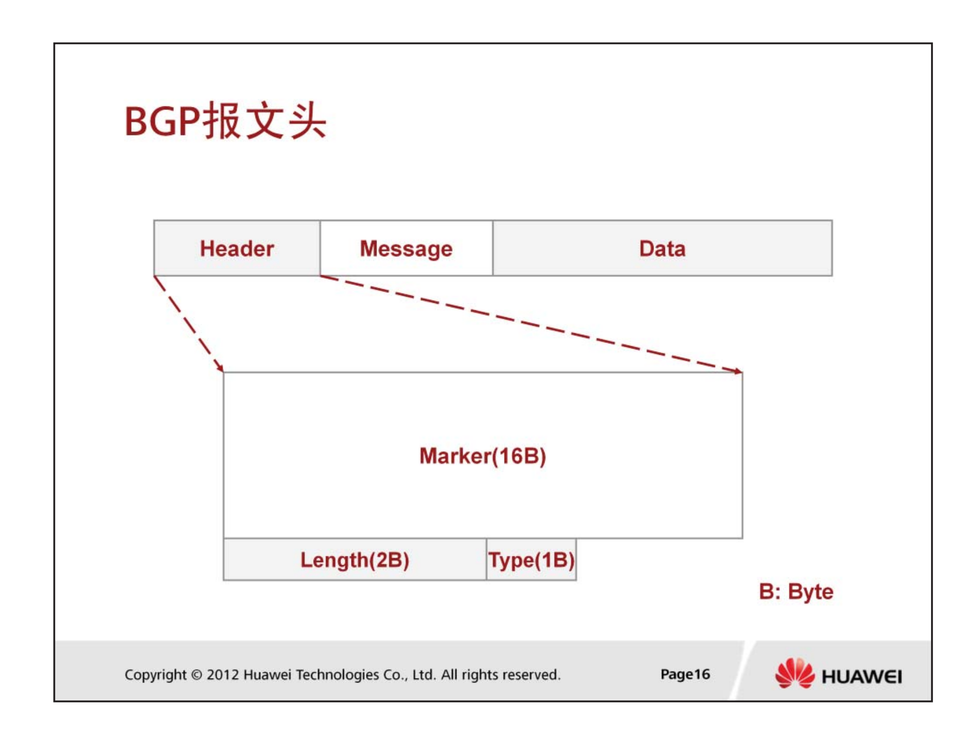
Keepalive：该报文用于BGP邻居关系的维护，为周期性交换的报文，用于判断对等体之间的可达性。

Update：该报文则是邻居之间用于交换路由信息的报文，其中包括撤销路由信息和可达路由信息及其各种路由属性。是BGP五个报文中最重要的报文。

Notification：BGP的差错检测机制，一旦检测到任何形式的差错，BGP Speaker会发送一个NOTIFICATION报文，随后与之相关的邻居关系将被关闭。

Route-refresh：用来通知对等体自己支持路由刷新能力；

其中，前四种消息是在RFC4271中定义的，而Refresh的消息则是在RFC2918中定义的 。



Marker（标记）：16字节，固定为1。

Length（长度）：两字节无符号整数。指定了消息的全长，包括头部。

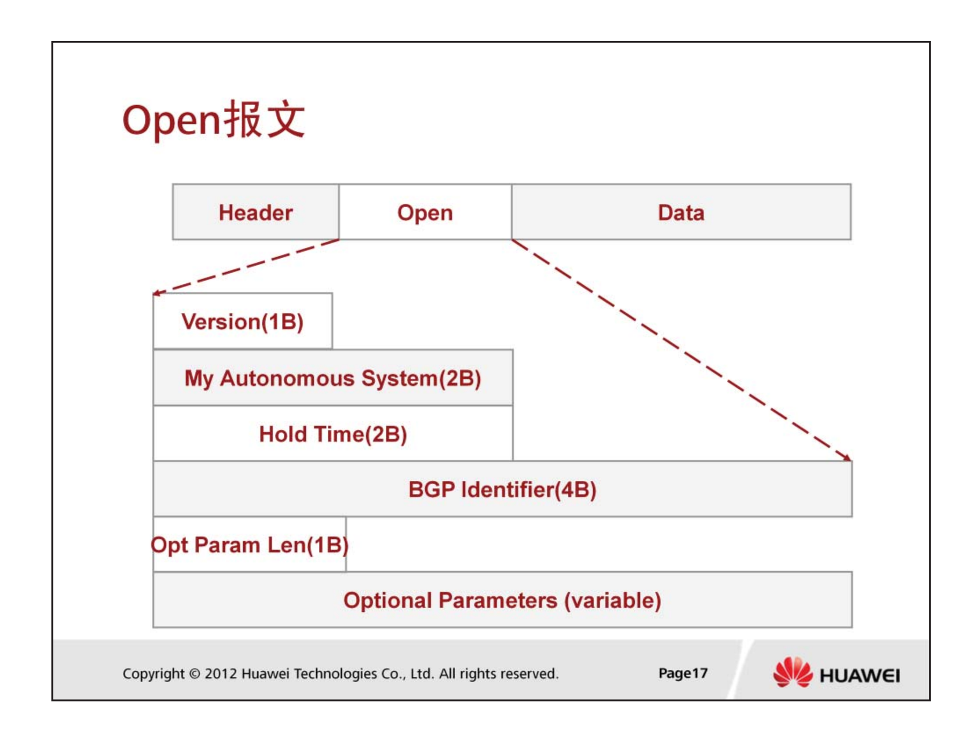
Type（类型）：1 字节，指示报文类型，如OPEN、UPDATE报文等 。

1 – OPEN

2 – UPDATE

3 – NOTIFICATION

4 – KEEPALIVE



主要字段的解释如下：

Version：BGP的版本号。对于BGPv4来说，其值为4。

My Autonomous System：本地AS编号。通过比较两端的AS编号可以确定是EBGP连接还是IBGP连接。

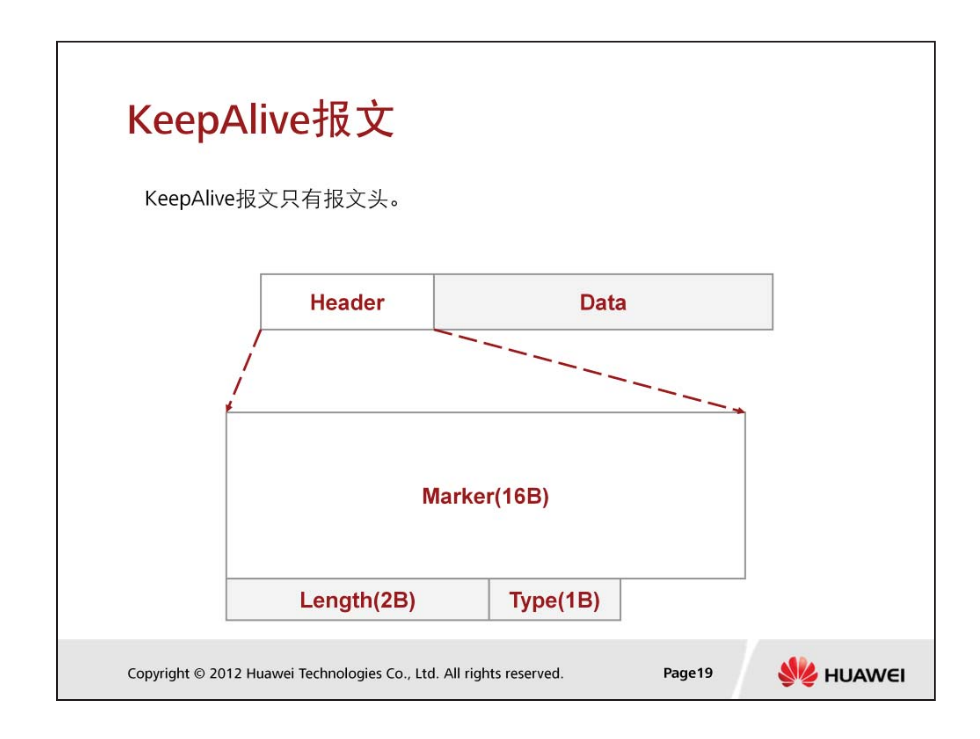
Hold Time：在建立对等体关系时两端要协商Hold time，并保持一致。如果两端所配置的Hold time时间不同，则BGP会选择较小的值作为协商的结果。如果在这个时间内未收到对端发来的Keepalive消息，则认为BGP连接中断。

BGP Identifier：BGP路由器的Router ID，以IP地址的形式表示，用来识别BGP路由器。在VRP5.30系统中，如果没有通过命令router id进行配置，则按照如下规则进行选择：优选Loopback接口地址中最大的地址作为Router ID，如果没有Loopback接口配置了IP地址，则从其它配置了IP地址的物理接口中选择一个最大IP地址的作为Router ID。

Opt Parm Len（Optional Parameters Length）：可选参数的长度。如果为0则没有可选参数。

Optional Parameters：是一个可选参数用于BGP验证或多协议扩展（Multiprotocol Extensions）等功能。每一个参数为一个（Parameter Type-Parameter Length-Parameter Value）三元组。

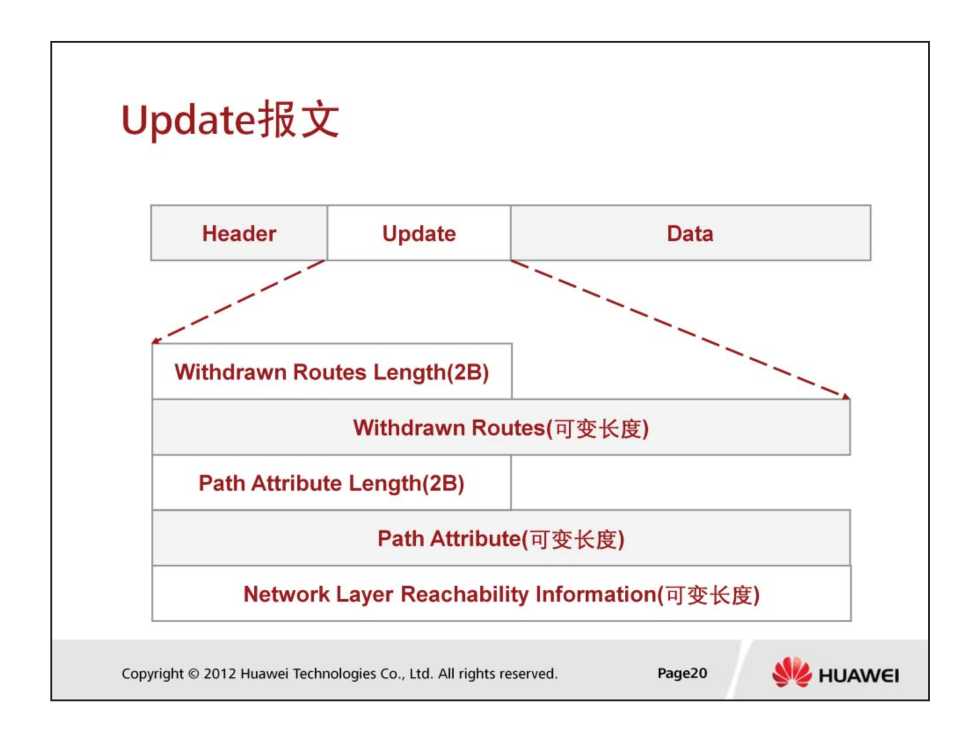
对等体在接收到Open消息后，将发送Keepalive消息确认并保持连接的有效性。确认后，对等体间可以进行Update、Notification、Keepalive和Route-refresh消息的交换。



KeepAlive报文主要用于对等体路由器间的运行状态以及链路的可用性确认。KeepAlive 报文的组成只包括一个BGP数据报头。 KeepAlive 消息在对等体之间的交换频率以保证对方保持定时器不超时为限。

当一台路由器与其邻居建立BGP连接之后，将以Keepalive-interval设定的时间间隔周期性地向对等体发送KeepAlive 报文，表明该连接是否还可保持。

缺省情况下，发送KeepAlive 的时间间隔为 60 秒，Hold Time是180秒。每次从邻居处接收到KeepAlive 报文将重置Hold Time定时器，如果Hold Time定时器超时，就认为对等体Down掉。



UPDATE消息被用作在BGP对等体之间传递路由信息。多条可达路由信息可以被通告到相应的对等体上，或者多条不可达路由信息被撤消。UPDATE消息由以下五个部分组成：

Withdrawn Routes Length：（2字节无符号整数）不可达路由长度，表示Withdrawn Routes字段的数据长度。如果Withdrawn Routes Length字段数值为0，则表示Withdrawn Routes字段没有任何数据，在UPDATE消息中不会被显示。

Withdrawn Routes：（变长）撤销路由。该字段包括一系列的IP地址前缀信息，以<length, prefix>的格式来表示，比如<19,198.18.160.0>表示一个198.18.160.0 255.255.224.0的网络。

Path Attribute Length：（2字节无符号整数）路由属性长度，表示Path Attribute字段的数据长度。如果Path Attribute Length数值为0，则表示Path Attribute字段没有任何数据，在UPDATE消息中不会被显示。

Path Attributes：（变长）路径属性。每个路径属性都是由三元组所组成：<attribute type, attribute length, attribute value>。

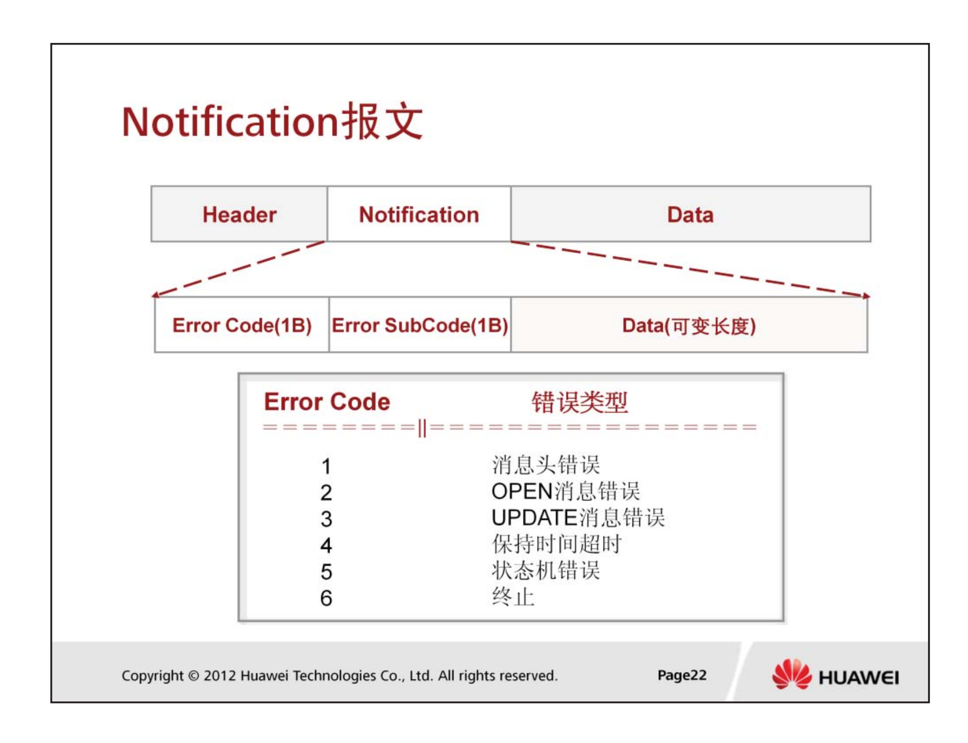
Network Layer Reachability Information：（变长）网络可达信息。包括一系列的IP地址前缀。格式与撤消路由字段一样<length, prefix>。

最小UPDATE消息的长度为23个字节(19字节的报文头+2字节的撤消路由长度+2字节的路径属性长度)。这样的UPDATE消息被称之为End-of-RIB，用于BGP GR。

一条UPDATE消息可以发布多条具有相同路由属性的可达路由，这些路由可共享一组路由属性。所有包含在一个给定的Update消息里的路由属性适用于该Update消息中的NLRI字段里的所有目的地（用IP前缀表示）。

一条UPDATE消息可以撤销多条不可达路由。每一个路由通过目的地（用IP前缀表示），清楚的定义了BGP Speaker之间先前通告过的路由。

一条UPDATE消息可以只用于撤销路由，这样就不需要包括路径属性或者网络可达信息。相反，也可以只用于通告可达路由，就不需要携带Withdrawn Routes了。



Notification报文主要在发生错误或对等体连接被关闭的情况下使用，该消息携带各种错误码（如定时器超时等），以及错误子码和错误信息。

Errorcode：错误码。1字节长的字段。每个不同的错误都使用唯一的代码表示，而每一个错误码都可以拥有一个或多个错误子码，但如果某些错误码并不存在错误子码的话，则该错误子码字段以全0表示。

Errsubcode：错误子码。

消息头错误子码：

1 – 连接非同步

2 – 错误的消息长度

3 – 错误的消息类型

OPEN消息错误子码：

1 – 不支持的版本号

2 – 错误的对等体AS号

3 – 错误的BGP ID

4 – 不支持的可选参数

5 – RFC1771里被定义为认证失败，

RFC4271里则对此表示反对。具体请参考RFC1771/RFC4271

6 – 不可接受的保持时间(Hold Time)

UPDATE消息错误子码：

1 – 畸形的属性列表

2 – 无法识别的公认属性

3 – 缺少的公认属性

4 – 属性标志位错误

5 – 属性长度错误

6 – 无效的ORIGIN属性

7 – RFC1771里被定义为AS路由环路，

RFC4271里对此表示反对。具体请参考RFC1771/RFC4271

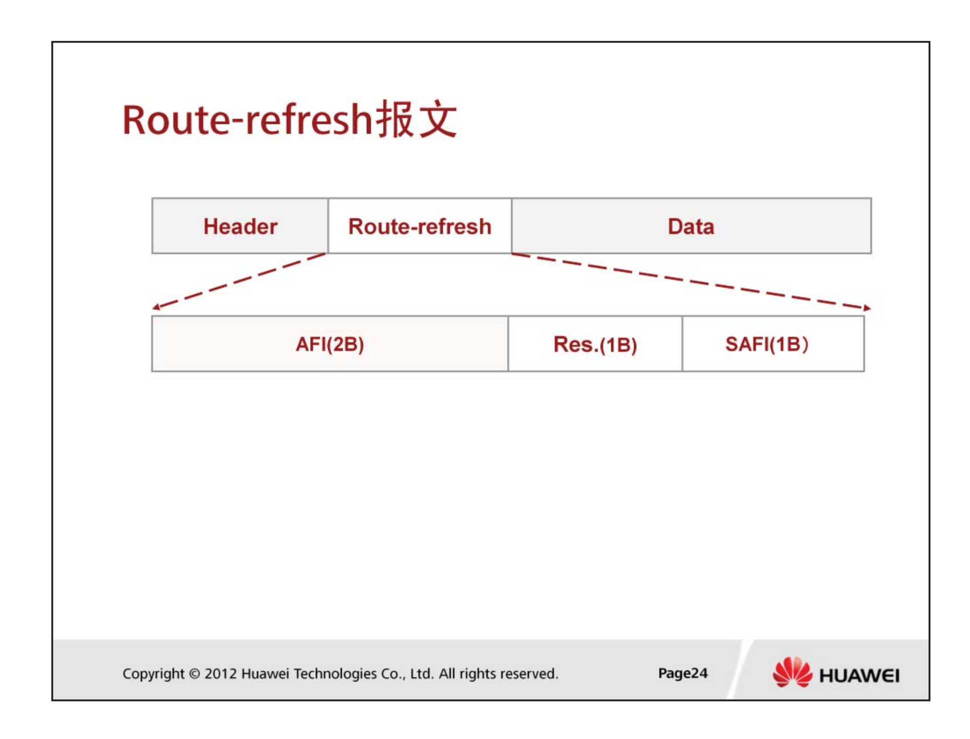
8 – 无效的下一跳属性

9 – 可选属性错误

10 – 无效的网络字段

11 – 畸形的AS_PATH

Data：依赖于不同的错误码和错误子码，用于标识错误原因。是一个可变长的字段，被NOTIFICATION用作诊断错误的原因。注：Data字段的长度可以由以下公式来决定：消息长度 = 21 + Data长度 （NOTIFICATION消息最小长度为21个字节，其中已经包括消息头。）



主要字段的解释如下:

AFI（Address Family Identifier）：地址族标识符（2字节）。

Res.（Reserved field）：保留区域（1字节），发送方应将其设置为0，接收方应当忽略该区域的信息。

SAFI（Subsequent Address Family Identifier）：子地址族标识符（8字节）。

在所有BGP路由器使能Route-refresh能力的情况下，如果BGP的入口路由策略发生了变化，本地BGP路由器会向对等体发布Route-refresh消息，收到此消息的对等体会将其路由信息重新发给本地BGP路由器。这样，可以在不中断BGP连接的情况下，对BGP路由表进行动态刷新，并应用新的路由策略。

BGP协议中消息的应用

- 通过TCP建立BGP连接时，发送OPEN消息
- 连接建立后，如果有路由需要发送或路由变化时，发送UPDATE消息通告对端
- 稳定后要定时发送KEEPALIVE消息以保持BGP连接的有效性
- 当本地BGP在运行中发现错误时，要发送NOTIFICATION消息通告BGP对等体
- ROUTE-REFRESH消息用来通知对等体自己支持路由刷新

 Page25



BGP使用TCP建立连接，本地监听端口为179。和TCP连接建立相同，BGP连接的建立也要经过一系列的对话和握手。TCP通过握手协商通告其端口等参数，BGP的握手协商的参数有：BGP版本、BGP连接保持时间、本地的路由器标识（Router ID）、授权信息等。这些信息都在Open消息中携带。

BGP连接建立后，如果有路由需要发送则发送Update消息通告对端。Update消息发布路由时，还要携带此路由的路由属性，用以帮助对端BGP协议选择最优路由。

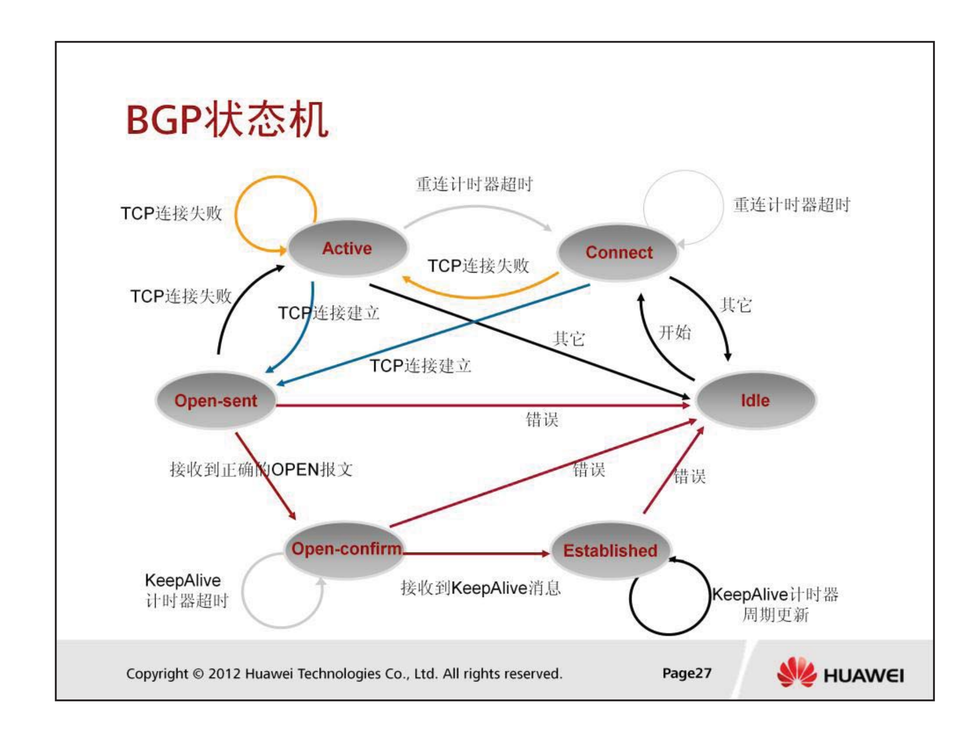
在本地BGP路由变化时，要通过Update消息来通知BGP对等体。

经过一段时间的路由信息交换后，本地BGP和对端BGP都无新路由通告，趋于稳定状态。此时要定时发送KEEPALIVE消息以保持BGP连接的有效性。对于本地BGP，如果在保持时间内，未收到任何对端发来的BGP消息，就认为此BGP连接已经中断，将断开此BGP连接，并删除所有从该对等体学来的BGP路由。

当本地BGP在运行中发现错误时，要发送NOTIFICATION消息通告BGP对

等体。如对端BGP版本本地不支持、本地BGP收到了结构非法的Update消息等。本地BGP退出BGP连接时也要发送NOTIFICATION消息。

ROUTE-REFRESH消息用来通知对等体自己支持路由刷新。



Idle：BGP连接的第一个状态。在空闲状态，BGP在等待一个启动事件。启动事件出现以后，BGP初始化资源，复位连接重试计时器（Connect-Retry），发起一条TCP连接，同时转入Connect（连接）状态。

Connect：在此状态，BGP发起第一个TCP连接，如果连接重试计时器超时，就重新发起TCP连接，并继续保持在Connect状态，如果TCP连接成功，就转入OpenSent状态，如果TCP连接失败，就转入Active状态。

Active：在此状态，BGP总是在试图建立TCP连接，如果连接重试计时器（Connect-Retry）超时，就退回到Connect状态，如果TCP连接成功，就转入OpenSent状态，如果TCP连接失败，就继续保持在Active状态，并继续发起TCP连接。

OpenSent：在此状态，TCP连接已经建立，BGP也已经发送了第一个Open报文，剩下的工作，BGP就在等待其对等体发送Open报文。并对收到的Open报文进行正确性检查，如果有错误，系统就会发送一条出错通知消息并退回到Idle状态，如果没有错误，BGP就开始发送Keepalive报文，并复位Keepalive计时器，开始计时。同时转入OpenConfirm状态。

OpenConfirm：在OpenConfirm状态，BGP等待一个Keepalive报文，同时复位保持计时器，如果收到了一个Keepalive报文，就转入Established阶段，BGP邻居关系就建立起来了。

Established：在Established状态，BGP邻居关系已经建立，这时，BGP将和它的邻居们交换Update报文，同时复位保持计时器。

另外，在除Idle状态以外的其它五个状态出现任何Error的时候，BGP状态机就会退回到Idle状态。

在BGP对等体建立的过程中，通常可见的三个状态是：Idle、Active、Established。

Idle状态下，BGP拒绝任何进入的连接请求，是BGP初始状态。

Active状态下，BGP将尝试进行TCP连接的建立，是BGP的中间状态。

Established状态下，BGP对等体间可以交换Update报文、Route-refresh报文、Keepalive报文和Notification报文。

BGP对等体双方的状态必须都为Established，BGP邻居关系才能成立，双方通过Update报文交换路由信息。



Page29



BGP数据库

IP路由表 (IP-RIB)

- 全局路由信息库，包括所有IP路由信息。

BGP路由表 (Loc-RIB)

- BGP路由信息库，包括本地BGP Speaker选择的路由信息。

邻居表

- 对等体邻居清单列表

Adj-RIB-In

- 对等体宣告给本地Speaker的未处理的路由信息库

Adj-RIB-Out

- 本地Speaker宣告给指定对等体的路由信息库

 Page30



IP路由表 (IP-RIB)

- 全局路由信息库，包括所有IP路由信息。

BGP路由表 (Loc-RIB)

- BGP路由信息库，包括本地BGP Speaker选择的路由信息。

邻居表

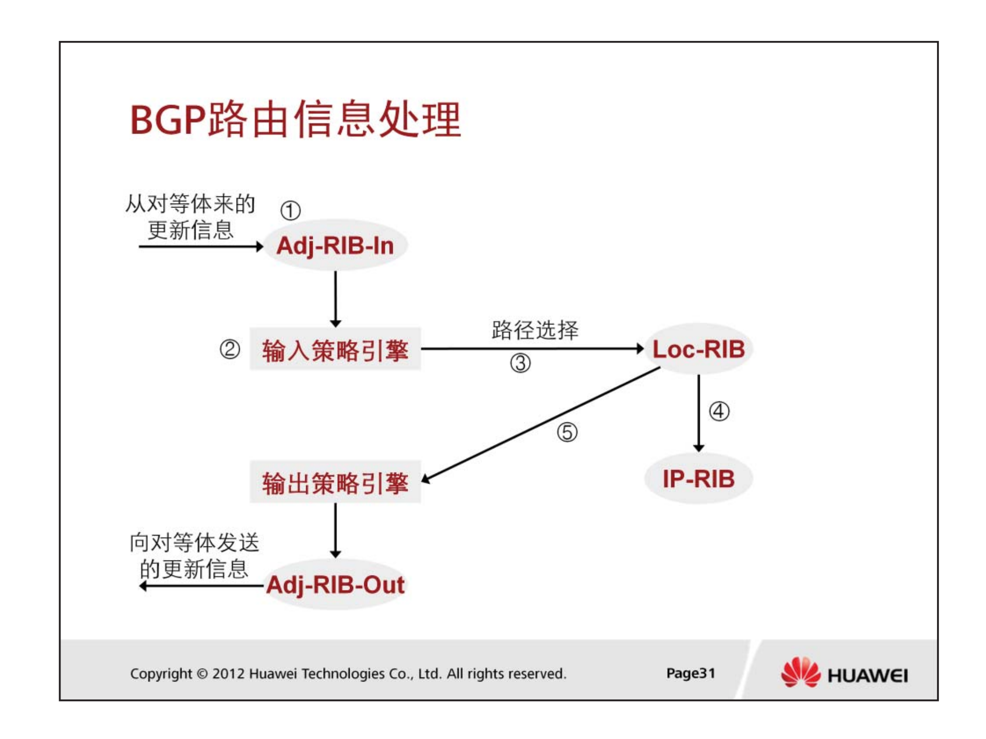
- 对等体邻居清单列表

Adj-RIB-In

- 对等体宣告给本地Speaker的未处理的路由信息库

Adj-RIB-Out

- 本地Speaker宣告给指定对等体的路由信息库



当从对等体接收到更新数据包时，路由器会把这些更新数据包存储到路由选择信息库(Routing Information Base, RIB)中，并指明是来自哪个对等体的(Adj-RIB-In)。这些更新数据包被输入策略引擎过滤后，路由器将会执行路径选择算法，来为每一条前缀确定最佳路径。

得出的最佳路径被存储到本地BGP RIB (Loc-RIB)中，然后被提交给本地IP路由选择表(IP-RIB)，以用作安装考虑。

如果启用了多路径特性，最佳路径和所有等值路径都将被提交给IP-RIB考虑。

除了从对等体接收来的最佳路径外，Loc-RIB也会包含当前路由器注入的(被称为本地发起的路由)，并被选择为最佳路径的BGP前缀。Loc-RIB中的内容在被通告给其他对等体之前，必须通过输出策略引擎。只有那些成功通过输出策略引擎的路由，才会被安装到输出RIB (Adj-RIB-Out)中。



问 题

BGP是怎样去发现邻居的?

BGP是基于什么传输层协议的?端口号为多少?

请说出BGP五种消息的作用?

 Page32



1.BGP是怎样去发现邻居的?

答:BGP并没有邻居发现机制,所有的邻居关系都是手工指定的。

2.BGP是基于什么传输层协议的?端口号为多少?

答:BGP是基于TCP协议的一个域间路由协议,其端口号为179。

3.请说出BGP五种消息的作用?

答:OPEN:主要包括BGP版本,AS号码等信息。试图建立BGP邻居关系的两个路由器建立了TCP会话后开始交换OPEN信息以最终确认能否形成邻居关系。

KEEPALIVE:该报文用于BGP邻居关系的维护,为周期性交换的报文,用于判断对等体之间的可达性。

NOTIFICATION:BGP的差错检测机制,一旦检测到任何形式的差错,BGP会发送一个NOTIFICATION报文,随后与之相关的邻居关系将被关闭。

UPDATE:这是BGP四个报文中最重要的报文。用于BGP邻居之间交换路由更新信息,它包括了BGP用来组建无环路的互连网络结构所需的所有信息,主要包括以下三个基本部分:

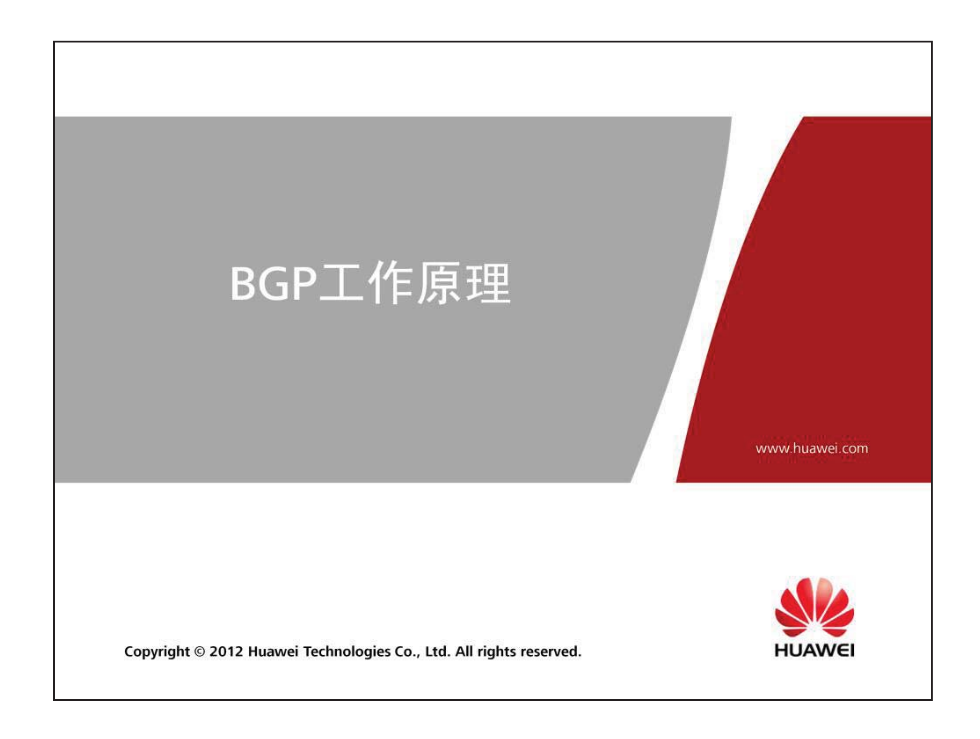
网络层可到达信息(NLRI)

路径属性（PATH ATTRIBUTES）

不可达的路由（WITHDRAWN ROUTES）

ROUTE-REFRESH：用来通知对等体自己支持路由刷新。


谢谢
www.huawei.com


BGP工作原理
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

BGP是主要工作在AS与AS间的动态路由协议，为AS间提供无环路的路由信息交互，而我们将会在接下来的胶片中介绍BGP到底如何提供AS间无环的路由信息交换。





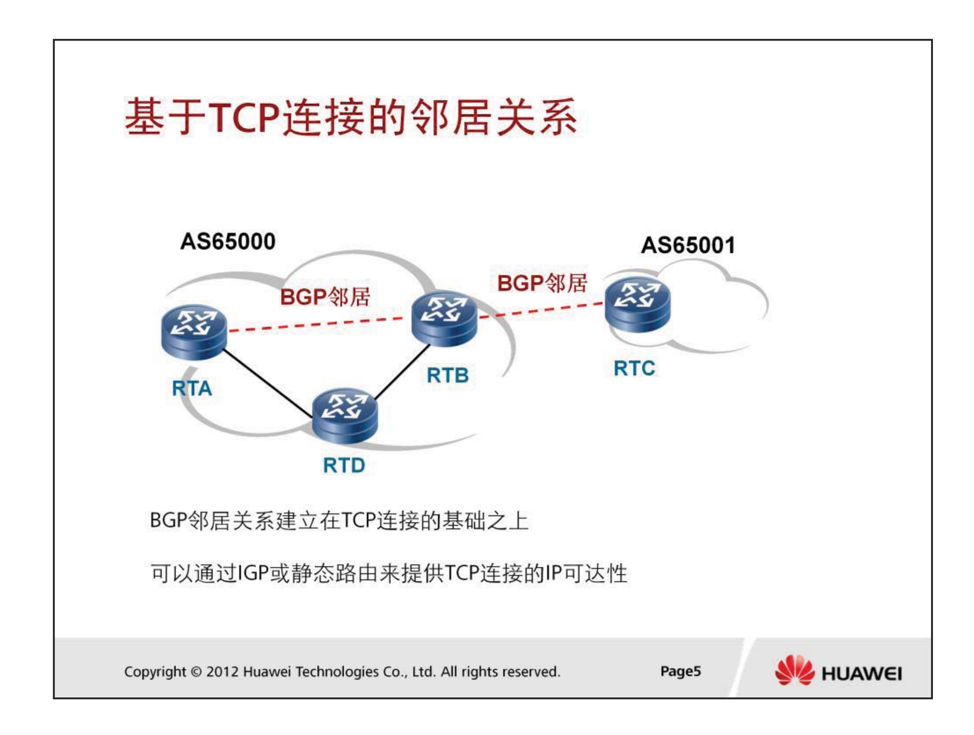
培训目标

学完本课程后，您应该能：

- 知道BGP的两种邻居关系
- 知道BGP的通告原则
- 知道BGP如何通告路由

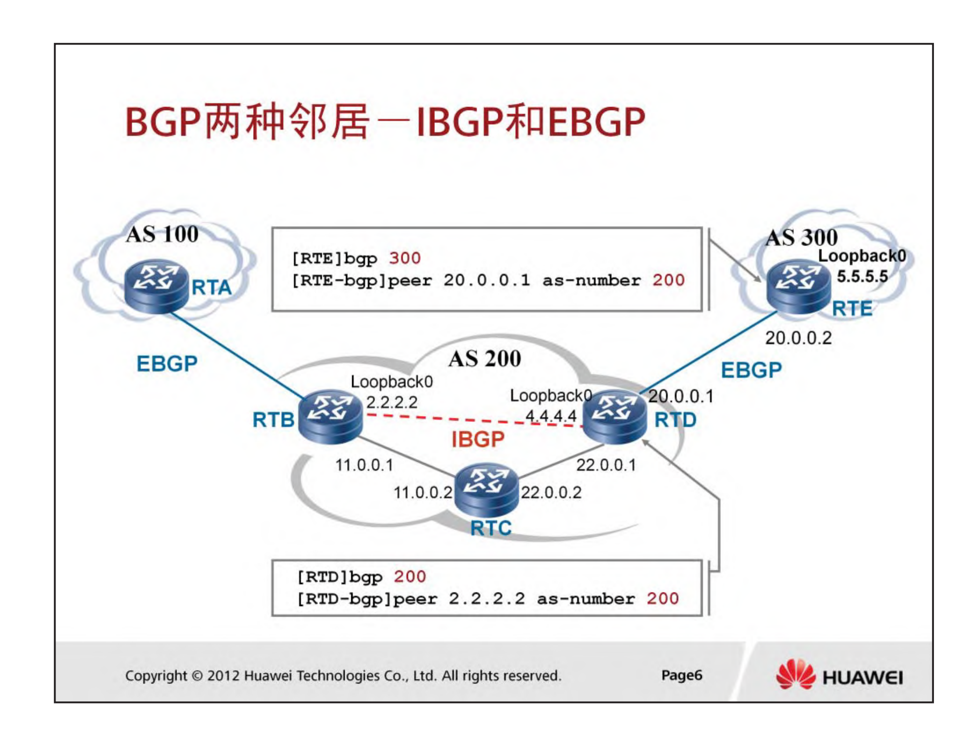



目 录
BGP邻居关系
BGP通告原则
BGP路由提高
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page4
HUAWEI



同OSPF、ISIS一样，在BGP中，路由学习的依然要首先建立邻居关系。所不同的是，OSPF、ISIS的邻居关系是自动建立的，而BGP邻居的建立必须手动完成，从邻居的建立开始就体现出了BGP是基于策略进行路由的（物理上直接相连未必是邻居，反过来物理上没有直接相连可以建立邻居关系）。

BGP邻居关系是建立在TCP会话的基础之上的，而两个运行BGP的路由器要建立TCP的会话就必须要具备IP连通性。IP连通性必须通过BGP之外的协议实现，具体来讲就是IP连通性通过内部网关协议（IGP）或者静态路由来实现，为方便起见，我们把通过内部网关协议或者静态路由实现的IP连通性统称为IGP 连通性或者IGP可达性（IGP Reachability）。



如果两个交换BGP报文的对等体属于同一个自治系统，那么这两个对等体就是IBGP对等体（Internal BGP），如RTB和RTD。如果两个交换BGP报文的对等体属于不同的自治系统，那么这两个对等体就是EBGP对等体（External BGP），如RTD和RTE。

虽然BGP是运行于自治系统之间的路由协议，但是一个AS的不同边界路由器之间也要建立BGP连接，只有这样才能实现路由信息在全网的传递，如RTB和RTD，为了建立AS100和AS300之间的通信，我们要在它们之间建立IBGP连接。

BGP的基本配置如下：

启动BGP（指定本地AS编号），进入BGP视图

[Router A] bgp as-number

bgp命令用来启动BGP，进入BGP视图，undo bgp命令用来关闭BGP。

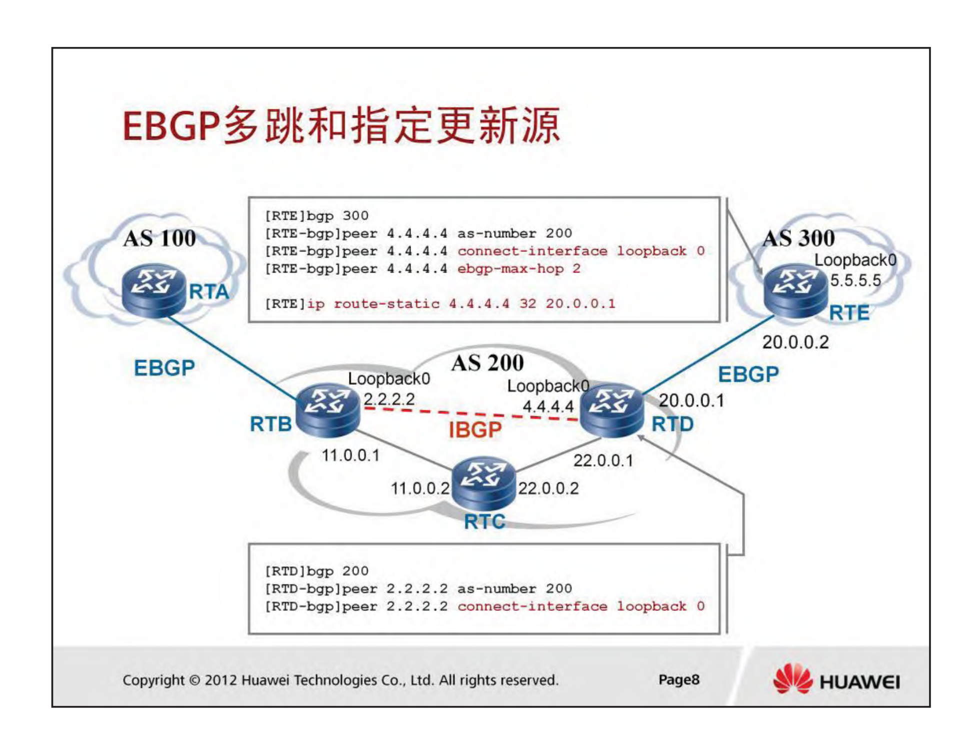
缺省情况下，系统不运行BGP。

每台路由器只能运行于一个AS内，即只能指定一个本地AS号。

指定对等体的IP地址及其所属的AS编号

[Router A-bgp] peer { group-name | ipv4-address | ipv6-address} as-number as-number

peer as-number命令用来配置指定对等体（组）的对端AS号，undo peer as-number命令用来删除对等体组的AS号。



IBGP对等体之间不一定是物理上直连的，只要TCP连接能够建立即可。为了IBGP对等体路由通告的可靠性，我们一般采用loopback接口建立IBGP邻居关系，在这种情况下，必须指定用于建立TCP连接的接口(也是路由更新报文的源接口)：

peer { group-name | peer-address } connect-interface interface-name

路由器一般默认要求EBGP对等体之间是有物理上的直连链路，同时一般也提供改变这个缺省设置的配置命令。允许同非直连相连网络上的邻居建立EBGP连接，这时需要修改EBGP报文的最大跳数：

peer { group-name | peer-address } ebgp-max-hop [ttl]


目 录
BGP邻居关系
BGP通告原则
BGP路由提高
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page9
HUAWEI

BGP路由通告原则(一)

连接一建立，BGP Speaker将把自己所有BGP路由通告给新对等体

多条路径时，BGP Speaker只选最优的给自己使用

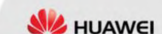
BGP Speaker只把自己使用的最优路由通告给对等体

```
[RTA]display bgp routing-table

 Total Number of Routes: 2

 BGP Local router ID is 1.1.1.1
 Status codes: * - valid, > - best, d - damped,
               h - history,  i - internal, s - suppressed, S - Stale
               Origin : i - IGP, e - EGP, ? - incomplete
      Network            NextHop        MED        LocPrf    PrefVal Path/Ogn

 *>i  192.168.3.0        10.1.1.2                             0      200i
 * i                     10.2.2.2                             0      200i
```

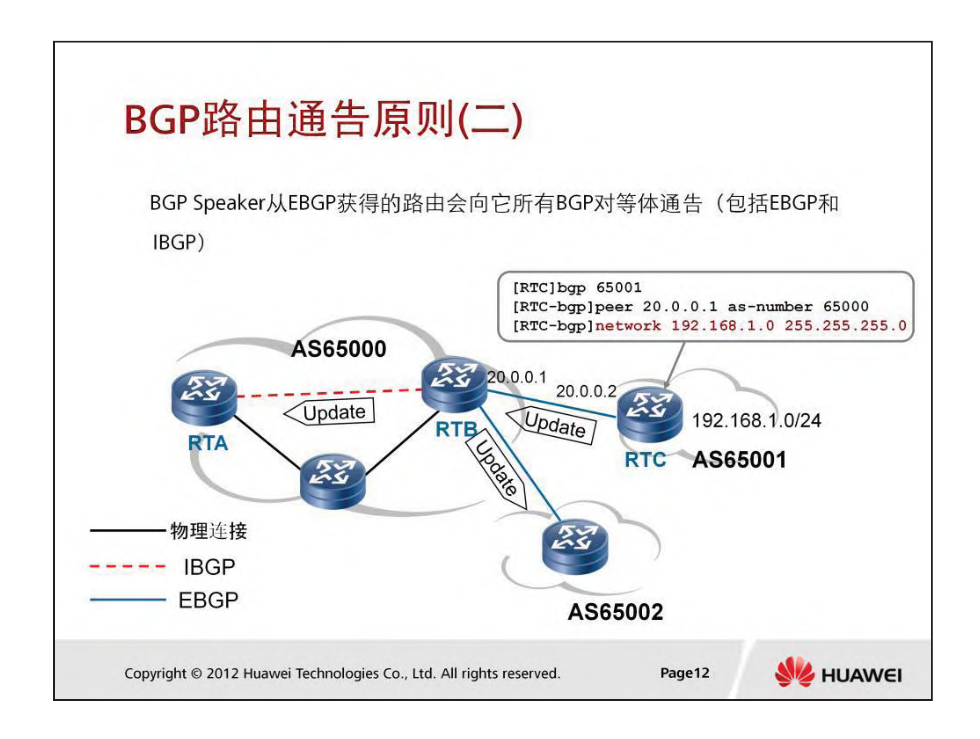
 Page10 

一般情况下，如果BGP Speaker学到去往同一网段的路由多于一条时，只会选择一条最优的路由给自己使用，即用来发布给邻居，同时上送给IP路由表。但是，由于路由器也会选择最优的路由给自己使用，所以BGP Speaker本身选择的最优的路由也不一定被路由器使用。例如，一条去往相同网段的BGP优选路由与一条静态路由，这时，由于BGP路由优先级要低，所以路由器会把这条静态路由加到路由表中去，而不会选择BGP优选的路由。

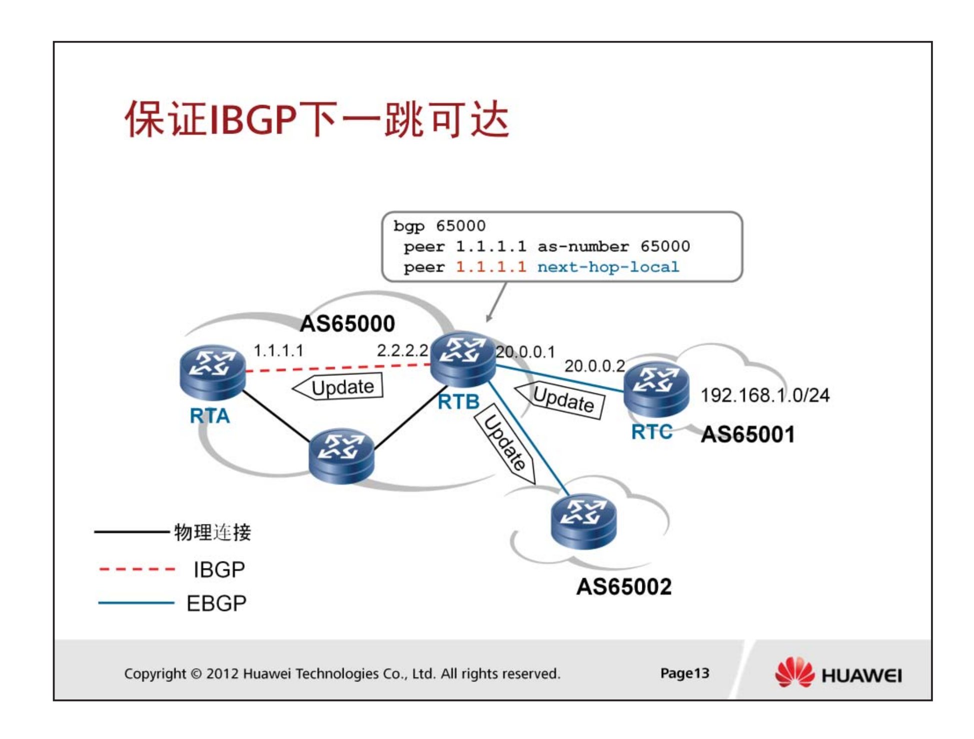
如胶片所示，当前RTA上存在两条去往192.168.3.0的路由，下一跳分别为10.1.1.2和10.2.2.2，BGP会根据选路原则选出最优路由(被打上“>”标记的路由)，用来发布给邻居。同时加入IP路由表，在IP路由表中会检查是否存在一条比BGP最佳路由更好的路由条目，比如有一条到达192.168.3.0的静态路由(静态路由的优先级为60，而BGP的优先级为255，数值越低越好)，则使用更优的路由条目，反之则把BGP最佳路由作为IP路由表的优选路由。

VRP5.7之前的版本，BGP优选同时也在IP路由表中优选的路由才能发布给邻居；

VRP5.7和之后的版本的缺省行为是BGP优选的路由即可发布给邻居，
同时提供一条命令active-route-advertise用于和之前的版本兼容



BGP Speaker从EBGP获得的路由会向它所有BGP对等体通告（包括EBGP和IBGP）。



对于IGP，工作原理是路由器之间交换路由信息，所以任何一个路由的下一跳是宣告此路由的路由器连接接口的IP地址，这是很容易理解的。而对于BGP，则主要是用于AS之间传递无环路的路由信息，BGP就是把AS抽象或者浓缩成一个路由器看待，所以RTB不会修改任何路由更新里的信息就更新给的RTA，即RTA要到达网络192.168.1.0/24，下一跳为20.0.0.2。这里又引入一个问题，对于RTA来说，很有可能不知道20.0.0.2的路由，这样就会导致路由不可达。

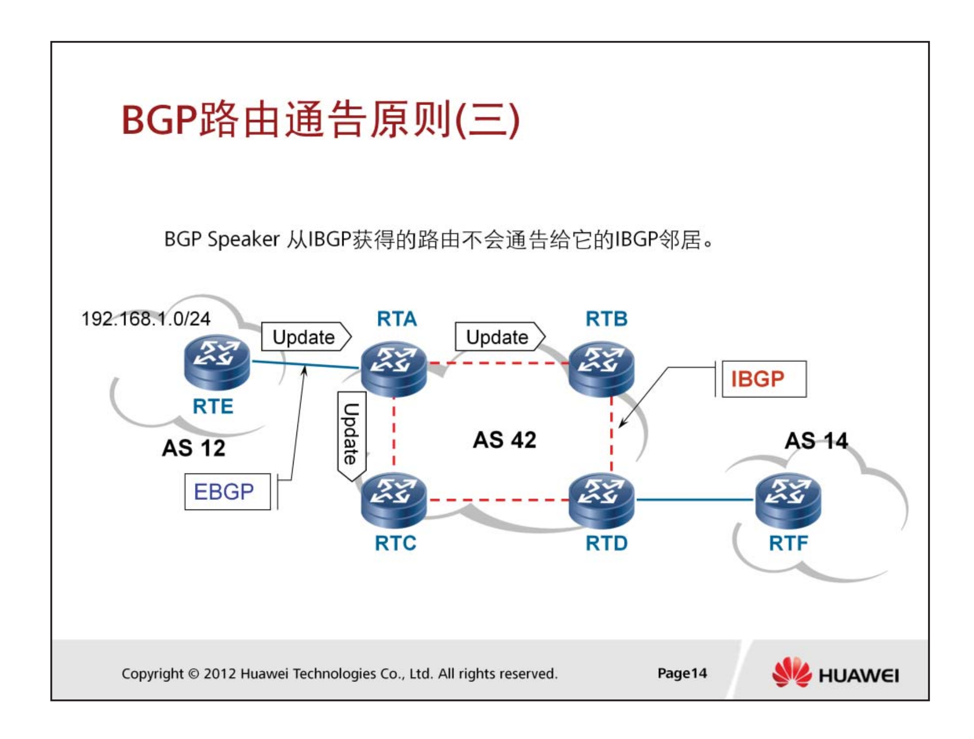
BGP提供了命令，让某些组网环境中，为保证IBGP邻居能够找到正确的下一跳，可以配置在向IBGP对等体发布路由时，改变下一跳地址为自身地址。

配置命令

peer { group-name | ipv4-address } next-hop-local

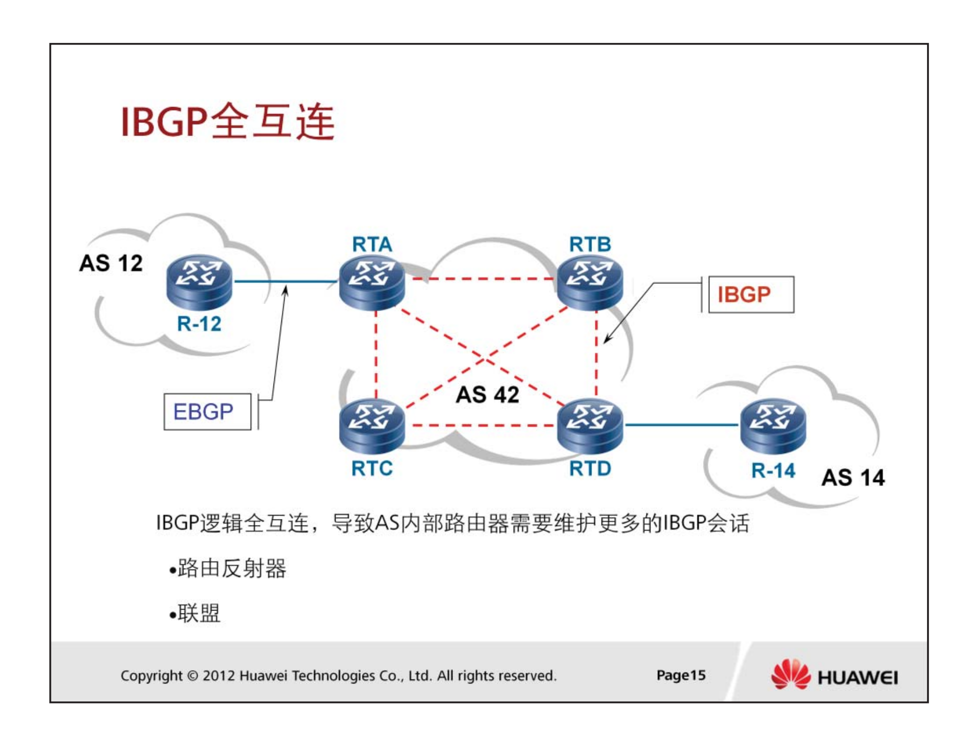
undo peer next-hop-local命令用来恢复缺省设置。

缺省情况下，BGP在向EBGP对等体通告路由时，将下一跳属性设为自身的IP地址。BGP在向IBGP对等体通告路由时，不改变下一跳属性。



如果没有这条路由通告规则，RTC从IBGP对等体RTA学到的路由就会通告给RTD，RTD继而会通告给RTB，RTB再把这条路由通告回RTA。这样就在AS内形成了路由环路。

所以，此原则是在AS内避免路由环路的重要手段。但是，这条原则的引入，带来了新的问题：RTD无法收到来自AS 12的BGP路由。一般我们采用IBGP的逻辑全连接来解决这个问题，即在RTA-RTD、RTB-RTC之间再建立两条IBGP连接。

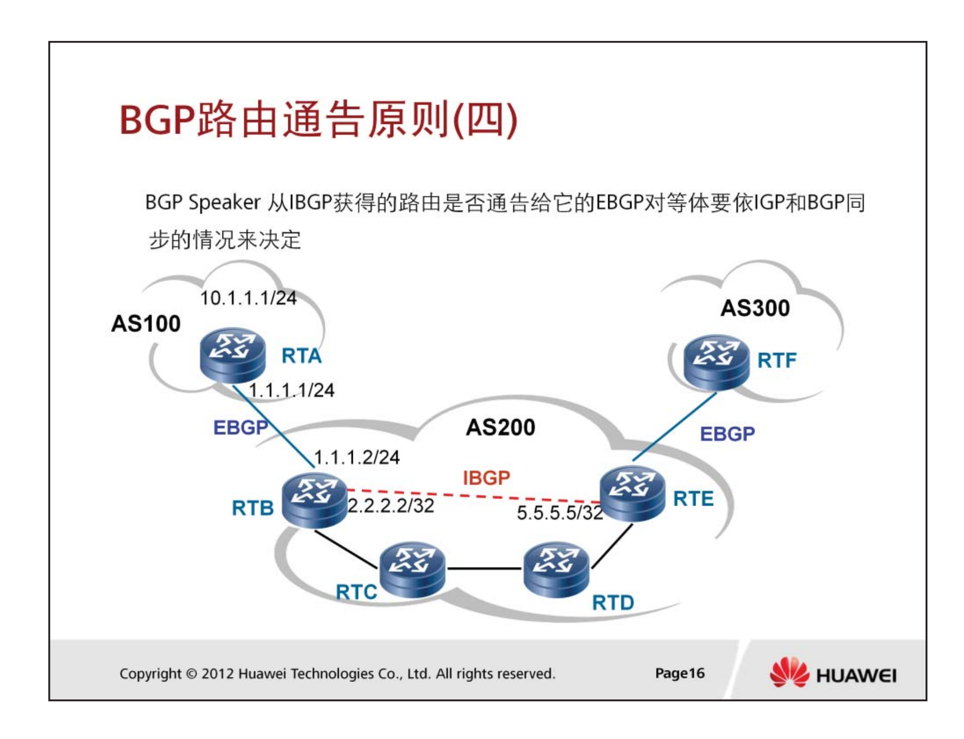


IBGP全互连（FULL-MESH）关系。这是解决由于IBGP水平分割带来的路由传递的问题的方法之一。这种方法的缺陷是路由器要付出更多的开销去维护网络里的IBGP会话。

除此以外，BGP还提供了如下两种解决IBGP水平分割的方案：

路由反射器（Route-Reflector）-- RFC 2796

联盟（Confederation）-- RFC 3065



BGP与IGP同步的概念：BGP Speaker不将从IBGP对等体获得的路由信息通告给它的EBGP对等体，除非该路由信息也能通过IGP获得。

BGP的主要任务之一就是向其它自治系统发布该自治系统的网络可达信息。如胶片所示，RTB会把去往10.1.1.0/24 的路由信息封装在BGP报文中，通过由RTB、RTE建立的TCP连接通告给RTE，如果RTE不考虑同步问题，直接接受了这条路由信息并通告给RTF。那么，如果RTF或RTE有去往10.1.1.0/24 的数据报文要发送，这个数据报文要想到达目的地必须经过RTD和RTC。但是，由于先前没有考虑同步问题，RTD和RTC的路由表中没有去往10.1.1.0/24的路由信息，数据报文到了RTD就会被丢弃。因此，BGP必须与IGP（如RIP、OSPF等）同步。也就是说，当一个路由器从IBGP对等体收到一条路由更新信息，在把它通告给它的EBGP对等体之前，要试图验证该目的地能否通过自治系统内部到达（即验证该目的地是否存在于IGP发现的路由表内，非BGP路由器是否可以传递报文到该目的地）。若能通过IGP知道这个目的地，才会把这样一条路由信息通告给EBGP对等体，否则认为BGP与IGP不同步，不进行通告。

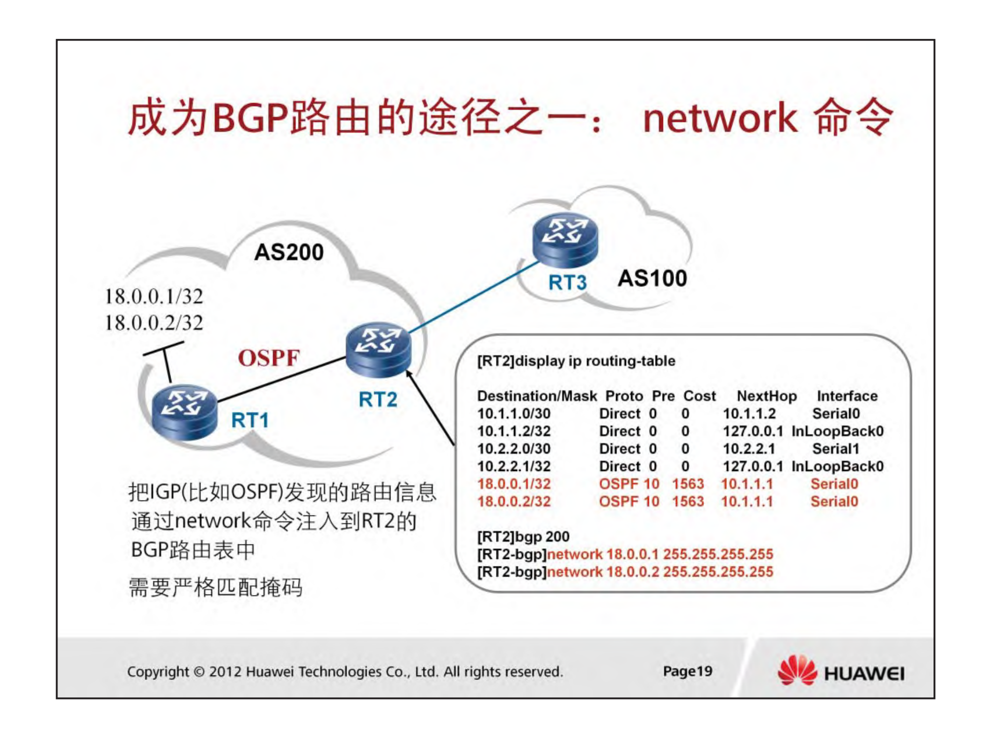
解决同步问题的方法有很多，最简单的办法是RTB把BGP路由信息引入到IGP中，这样就同步了。但是一般不建议这样做，因为BGP路由

表很大，引入到IGP中来会给系统带来很大负担，甚至导致中低端路由设备的瘫痪。其它的解决办法如：可以在RTB上配置一条去往10.1.1.0/24的静态路由，再把该静态路由引入到IGP中，这样也可以达到同步。但不论何种方法，都不适用于大规模网络。

实际上，VRP平台缺省情况下BGP与IGP是非同步的，并不可改变。

但取消同步是有条件的。当AS中所有的BGP路由器能组成IBGP全闭合网时，才可以取消同步，即RTB-RTC、RTB-RTD、RTB-RTE、RTC-RTD、RTC-RTE、RTD-RTE都通过TCP连接建立IBGP邻居关系。这时，我们来看，数据到RTD后，由于RTB-RTD建立了IBGP邻居，所以RTD上有去往10.1.1.0/24的从RTB学来的BGP路由，这时，通过路由迭代，RTD将数据发给RTC；同理，RTC也会把数据发给RTB。这样，数据就不会在途中丢失了。


目 录
BGP邻居关系
BGP通告原则
BGP路由提高
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page18
HUAWEI



BGP的主要工作是在自治系统之间传递路由信息，而不是去发现和计算路由信息。所以，路由信息需要通过配置命令的方式注入到BGP中。

成为BGP路由有两种配置方法：通过Network命令以及通过Import命令。

另外用aggregate命令也可以把路由注入到BGP路由表中，但由于aggregate注入的条件为：BGP路由表里已经明确存在明细路由的情况下，才能通过aggregate命令注入聚合路由，所以在这里不被归纳为其中方法之一。

通过Network命令：路由器将通过Network将IP路由表里的路由信息注入到BGP的路由表中，并通过BGP传递给其它对等体。通过Network命令注入到BGP路由表里的路由信息必须存在于IP路由表中。

相关命令：

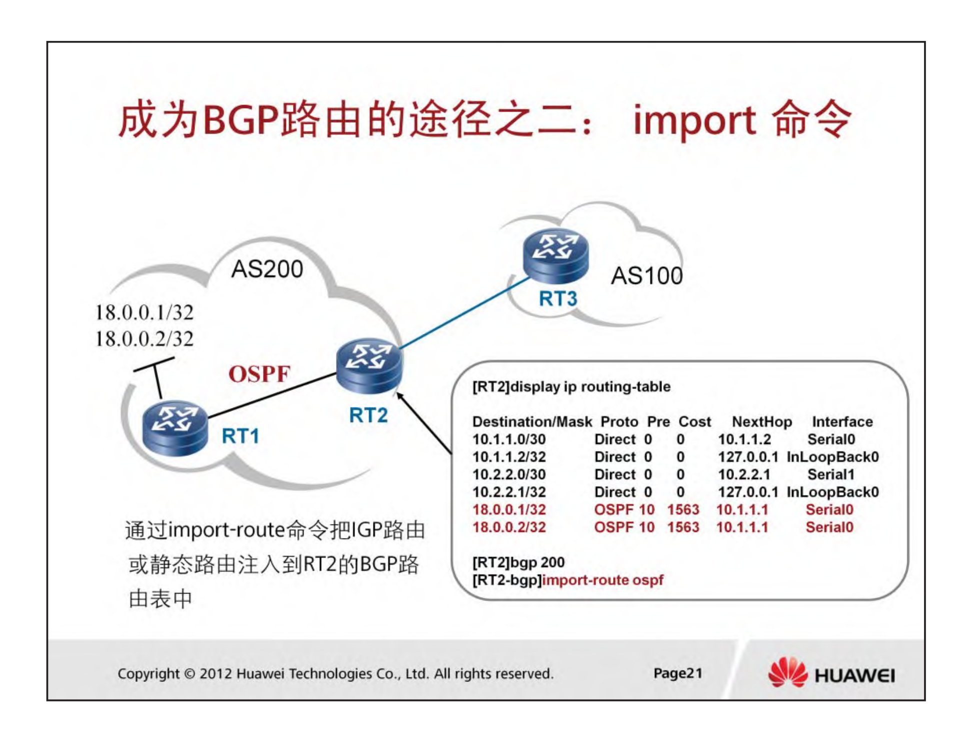
network ipv4-address [mask | mask-length] [route-policy route-policy-name]

ipv4-address：BGP发布的IPv4网络地址，点分十进制形式。

mask/mask-length：IP地址掩码或掩码长度。如果没有指定掩码，则按有类地址处理。

route-policy-name：发布路由应用的路由策略。

缺省情况下，BGP不发布任何本地的网络路由。



Destination/Mask	Proto	Pre	Cost	NextHop	Interface
10.1.1.0/30	Direct	0	0	10.1.1.2	Serial0
10.1.1.2/32	Direct	0	0	127.0.0.1	InLoopBack0
10.2.2.0/30	Direct	0	0	10.2.2.1	Serial1
10.2.2.1/32	Direct	0	0	127.0.0.1	InLoopBack0
18.0.0.1/32	OSPF	10	1563	10.1.1.1	Serial0
18.0.0.2/32	OSPF	10	1563	10.1.1.1	Serial0

第二种方法是通过Import命令把其它协议的路由信息注入到BGP路由表中，通过Import注入的路由信息可以结合策略共同使用。

import-route protocol [process-id] [med med | route-policy route-policy-name]

protocol：指定可引入的外部路由协议，目前包括isis、ospf、static 、direct和rip。

process-id：当引入路由协议为isis、ospf或rip时，必须指定进程号。

med：指定引入路由的MED度量值，取值范围0～65535。

route-policy-name：从其他路由协议引入路由时，可以使用该参数指定的路由策略过滤路由。





1.BGP的邻居关系有多少种?

答：IBGP和EBGP。如果两个交换BGP报文的对等体属于同一个自治系统，那么这两个对等体就是IBGP对等体（Internal BGP）。如果两个交换BGP报文的对等体属于不同的自治系统，那么这两个对等体就是EBGP对等体（External BGP）。

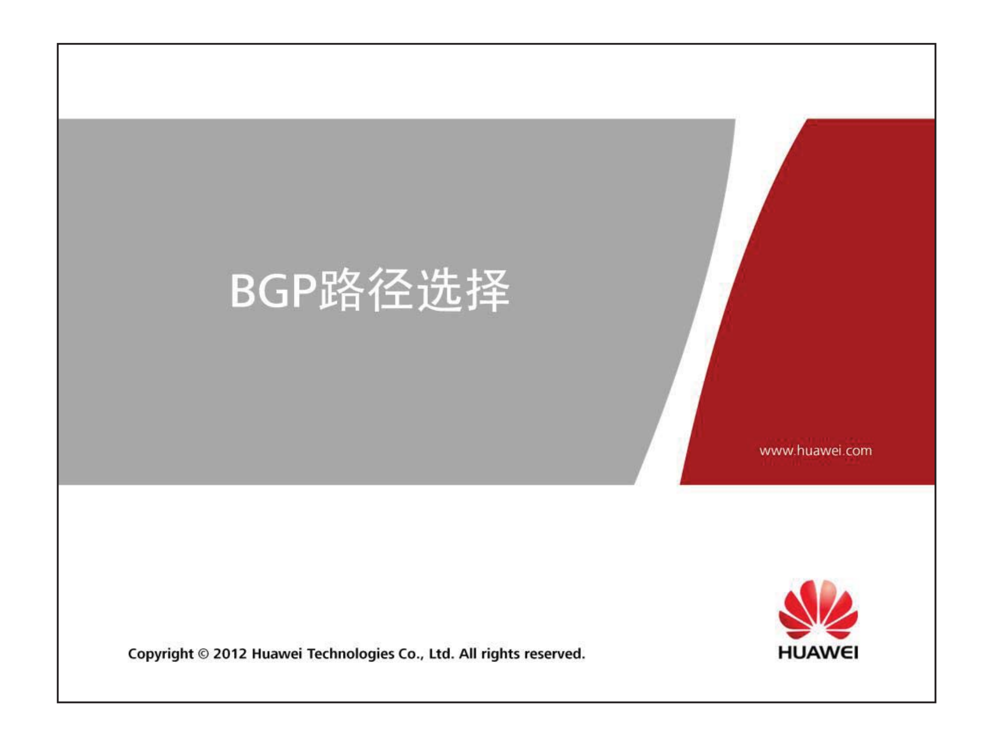
2.AS内部的IBGP对等体为什么需要建立全互连?

答：为防止AS内部的路由环路，BGP Speaker从IBGP获得的路由不会通告给它的IBGP邻居。在默认的情况下，为了能将路由信息成功传递到其它的IBGP对等体，BGP Speaker必须与其它BGP对等体都建立IBGP连接，在AS内部形成IBGP全互连。

3.通过network命令将路由注入到BGP中，需要什么条件?

答：通过network注入的路由必须存在于IP路由表中，而且注入的路由需要严格匹配IP路由表中的掩码长度。


谢谢
www.huawei.com


BGP路径选择
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

BGP作为一个策略工具，主要作用是实现AS间的路由信息传递。BGP就是结合丰富的路径属性，很好的控制路由信息的传递，从而实现路径的选择。





学完本课程后，您应该能：

- 知道什么是路径属性
- 了解BGP常用路径属性
- 了解BGP的选路原则

Page2




目 录
BGP路径属性介绍
BGP路径选择
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page4
HUAWEI

BGP的路径属性

BGP路径属性是一组描述BGP前缀特性的参数

BGP路径属性可以被分为四大类:

- 公认必遵 (Well-known mandatory)
- 公认任意 (Well-known discretionary)
- 可选过渡 (Optional transitive)
- 可选非过渡 (Optional non-transitive)

 Page5



对于企业和服务供应商所关心的问题，如：如何过滤某些BGP路由？如何影响BGP的选路？通过使用BGP丰富的路由属性，就可以得到解决。

BGP路由属性是一套参数，它对特定的路由进行更详细的描述。在配置路由策略时我们将广泛地使用各种路由属性。

BGP路径属性可以被分为四大类:

- 公认必遵 (Well-known mandatory)
- 公认任意 (Well-known discretionary)
- 可选过渡 (Optional transitive)
- 可选非过渡 (Optional non-transitive)

BGP的路径属性 (续)

公认属性是所有BGP路由器都必须识别的属性

- 公认必遵 (Well-known mandatory)
  - 所有BGP路由器都可以识别，且必须存在于Update消息中。如果缺少这种属性，路由信息就会出错
- 公认任意 (Well-known discretionary)
  - 所有BGP路由器都可以识别，但不要求必须存在于Update消息中，可以根据具体情况来决定是否添加到Update消息中

Page6



BGP必须识别所有公认属性。而一些强制属性必须包含在每一个UPDATE消息里，而其它任意属性则可能会被包含在某具体UPDATE消息中。一旦BGP对等体更新带有公认属性的UPDATE消息时，BGP对等体必须转发这些公认属性给其它对等体。

公认属性是所有BGP路由器都必须识别的属性：

- 公认必遵 (Well-known mandatory)
  - 所有BGP路由器都可以识别，且必须存在于Update消息中。如果缺少这种属性，路由信息就会出错
- 公认任意 (Well-known discretionary)
  - 所有BGP路由器都可以识别，但不要求必须存在于Update消息中，可以根据具体情况来决定是否添加到Update消息中

BGP的路径属性 (续)

可选属性不需要都被BGP路由器所识别

- 可选过渡 (Optional transitive)
  - BGP路由器可以选择是否在Update消息中携带这种属性。接收的路由器如果不识别这种属性，可以转发给邻居路由器，邻居路由器可能会识别并使用到这种属性
- 可选非过渡 (Optional non-transitive)
  - BGP路由器可以选择是否在Update消息中携带这种属性。在整个路由发布的路径上，如果部分路由器不能识别这种属性，可能会导致该属性无法发挥效用。因此接收的路由器如果不识别这种属性，将丢弃这种属性，不必再转发给邻居路由器

 Page7 

除公认属性外，每UPDATE消息里都可以包含一个或多个可选属性。并且不是每个BGP Speaker都要求支持这些可选属性。

而一个新的可过渡属性可以被发起者或其它一些BGP Speaker添加到路径属性上。

可选属性不需要都被BGP路由器所识别：

- 可选过渡 (Optional transitive)
  - BGP路由器可以选择是否在Update消息中携带这种属性。接收的路由器如果不识别这种属性，可以转发给邻居路由器，邻居路由器可能会识别并使用到这种属性

- 可选非过渡 (Optional non-transitive)
  - BGP路由器可以选择是否在Update消息中携带这种属性。在整个路由发布的路径上，如果部分路由器不能识别这种属性，可能会导致该属性无法发挥效用。因此接收的路由器如果不识别这种属性，将

丢弃这种属性，不必再转发给邻居路由器。

常见BGP路由属性

1、Origin
2、AS_PATH
3、Next hop
4、MED
5、Local-Preference
6、Atomic-Aggregate
7、Aggregator
8、Community
9、Originator-ID
10、Cluster-List
11、MP_Reach_NLRI
12、MP_Unreach_NLRI
13、Extended_Communities

 Page9 

以下列出几种常用的属性:

Origin：起点属性。定义路由信息的来源，标记一条路由是怎样成为BGP路由的。

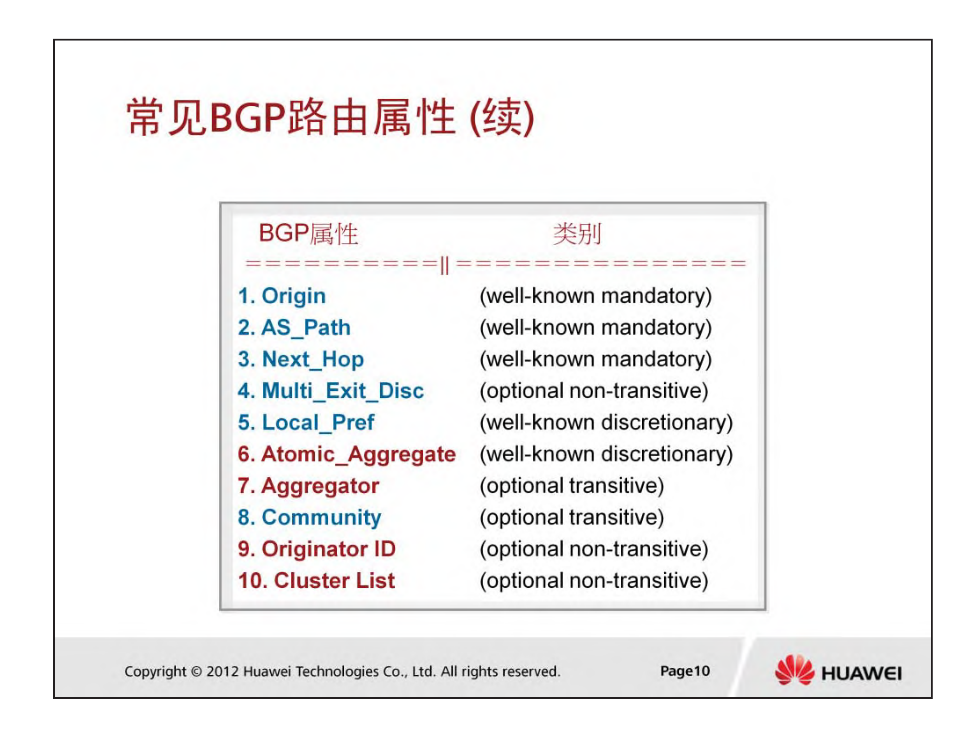
As_PATH：AS路径属性。是路由经过的AS的序列，即列出此路由在传递过程中经过了哪些AS。它可以防止路由循环，并用于路由的过滤和选择。

Next hop：下一跳属性。包含到达更新消息所列网络的下一跳边界路由器的IP地址。

MED属性：当某个AS有多个入口时，可以用MED属性来帮助其外部的AS选择一个较好的入口路径。一条路由的MED值越小，其优先级越高。

Local-Preference：本地优先级属性。用于在AS内优选到达某一目的地的路由。反映了BGP Speaker对每条BGP路由的偏好程度。属性值越大越优。

Community：团体属性。团体属性标识了一组具有相同特征的路由信息，与它所在的IP子网或自治系统无关。


常见BGP路由属性 (续)
BGP属性 类别
1. Origin (well-known mandatory)
2. AS_Path (well-known mandatory)
3. Next_Hop (well-known mandatory)
4. Multi_Exit_Disc (optional non-transitive)
5. Local_Pref (well-known discretionary)
6. Atomic_Aggregate (well-known discretionary)
7. Aggregator (optional transitive)
8. Community (optional transitive)
9. Originator ID (optional non-transitive)
10. Cluster List (optional non-transitive)
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page10
HUAWEI

起源 (Origin) 属性

一般的，具体的实现按如下方式决定一条路由的Origin属性

- 某条路由是直接而具体的注入到BGP路由表中的，则origin属性为IGP
  - 通过network命令注入BGP的路由
- 通过EGP(RFC904)学到的路由，则origin属性为EGP
- 其他情形下，Origin属性都为 Incomplete
  - 通过import命令注入BGP的路由

Origin属性值默认情况下不被任何路由器修改

 Page11



该属性定义了BGP路径信息源头，实际上也就是BGP speaker产生BGP路由的方式，有如下三个值：

IGP：在BGP路由表中（用display bgp routing-table查看）将会看到“i”的标识。通过network命令宣告的路由，起点属性为IGP，此种方式也称为BGP信息的半动态注入。 network命令所宣告的网络来自于IGP协议（包括静态路由），这些路由是有选择性的通过network命令转换为BGP路由，所以称为“半动态”。

EGP：在BGP路由表中将会看到“E”的标识，通过EGP转化(import)的BGP路由将具备此属性，这个属性我们在现实网络中将很难遇到，因为EGP这个协议基本上已经退出了历史舞台。

Incomplete：在BGP路由表中将会有一个“？”标识，具备这种属性的路由是通过一些别的方式学到的，属于未知的不明确的状态。一般来说，是通过将IGP或者静态路由引入（import）以后产生的。因为无条件的把IGP信息引入到BGP可能会造成副作用——不要的或者错误的信息会泄露（leak）进BGP中，比如IGP中可能会包含很多仅仅用于AS内部的专用地址或者未经注册的地址。除此以外，这样做还有可能造成BGP的动荡（因为BGP的路由依赖于IGP路由），对此问题BGP提供了一个解决方案，路由衰减（ROUTE DAMPENING），此处我们将不再讨论。

在这种情况下，我们必须要施加特殊的过滤，以确定哪些特定

的网络可以从IGP注入到BGP中。对于能区分开内部和外部路由的协议，比如OSPF，我们可以通过配置来保证仅仅将内部路由注入到BGP中（VRP5中默认情况只会引入OSPF内部路由，并不会引入OSPF外部路由到BGP中）；BGP的路由还可以通过引入静态路由并下发，这样做可以提高路由的稳定性。

起点属性三个值的优先顺序为IGP>EGP>INCOMPLETE，这三个值对于BGP的选路起着控制作用。

起源 (Origin) 属性 (续)

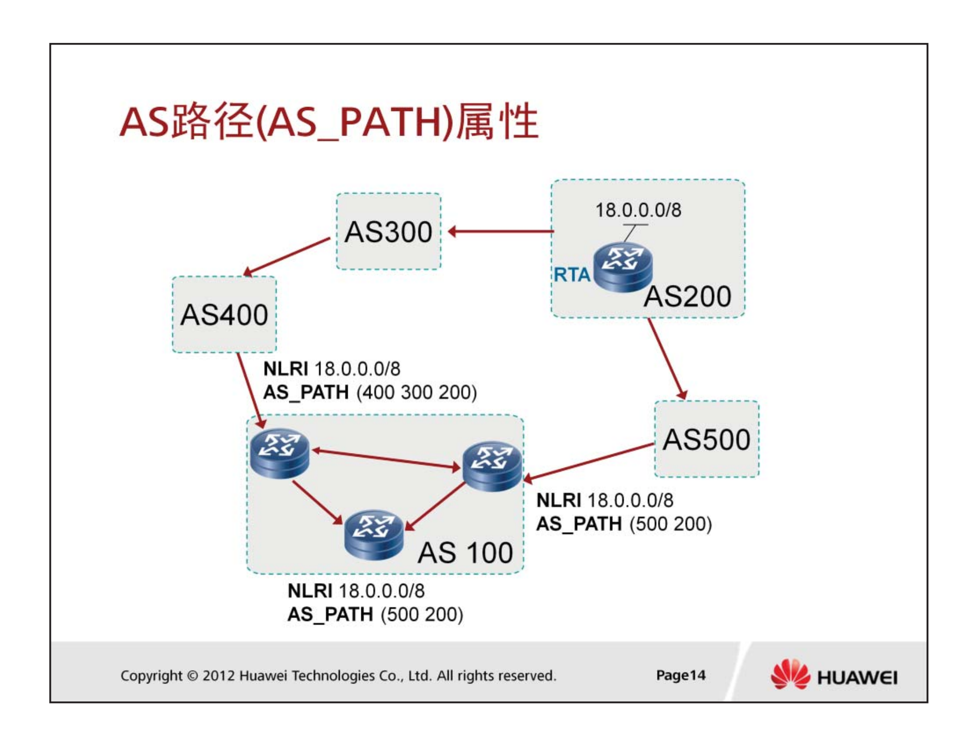
```
[RTB]display bgp routing-table

 Total Number of Routes: 2

 BGP Local router ID is 192.168.2.1
 Status codes: * - valid, > - best, d - damped,
               h - history,  i - internal, s - suppressed, S - Stale
               Origin : i - IGP, e - EGP, ? - incomplete
      Network            NextHop        MED        LocPrf    PrefVal Path/Ogn

 *>   192.168.1.0        10.1.1.1       0                     0       100i
 *    192.168.2.0        10.1.1.1       0                     0       100i
```



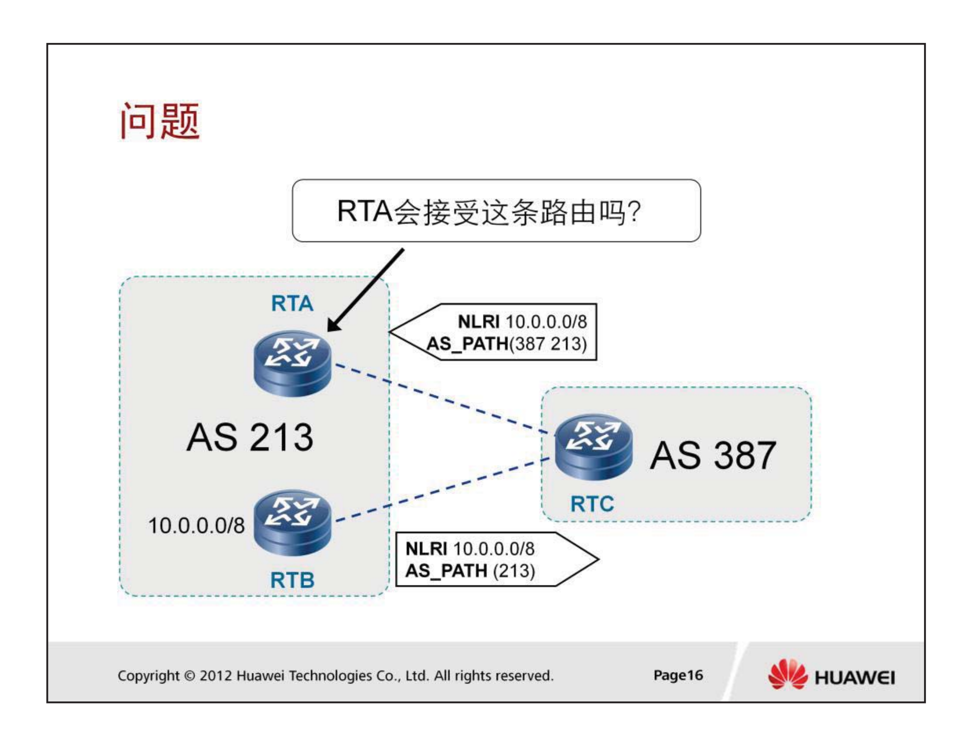
所谓AS_PATH是指BGP路由在传输的路径中所经历的AS的列表，是BGP中一个非常重要的公认必遵属性。 BGP不会接受AS_PATH属性中包含本AS Number的路由，从而避免了产生环路的可能。为此，BGP在向EBGP对等体通告一条路由时，要把自己的AS号加入到AS_PATH属性中，以记录此路由经过AS的信息，如果在路由更新消息中发现自己所在的AS号已经被包含在AS_PATH属性中，则表明该路由之前曾经通过该AS或者是源自于该AS，为避免路由环路，应该将此路由信息丢弃。

另外，AS_PATH属性在路径选择上也是一个很重要的衡量参数。当路由器中存在两条或者两条以上的到同一目的地的路由时，这些路由可以通过此属性比较相互之间的优劣，AS_PATH越短的路径越优先。注意：在大多数的实际网络中，多条路径的优劣往往是由AS_PATH来决定。

如胶片所示，AS200内的关于网络18.0.0.0/8的BGP路由经AS200、AS300、AS400到达AS100的AS_PATH为（400 300 200），经AS200、AS500到达AS100的AS_PATH为（500 200），这时BGP优先选择有较短AS_PATH的BGP路由：“500 200”。

在进行BGP的路由聚合时，缺省情况下形成聚合路由的具体路由其独特性将会丢失。这样一来如果某一个AS将来自于不同的其他AS的具体路由聚合。聚合路由的AS_PATH中将不会包含具体路由的AS号，这样此聚合路由就有可能会传回到其具体路由所在的AS中，从而形成路由环路。对此问题我们还将在路由聚合一节进行深入的讨论。

另外在进行路由过滤时，基于AS_PATH的过滤列表在很多情况下相对于一般的前缀列表能提供更加灵活的控制。



在默认情况下，BGP是通过AS号来检测路由环路的。如胶片所示，RTA-RTC、RTB-RTC建立EBGP邻居关系，当RTB将路由信息通告给RTC时带上本AS号（213）。然后RTA再从RTC接收路由时，路由的AS_PATH属性中就会带有本AS号（213），所以RTA不会接受这条路由信息。

但是在某些特殊应用中，如Hub&Spoke组网方式下，我们需要接受AS号重复的BGP路由。此时，可以用下面的命令来强制接受此类路由：

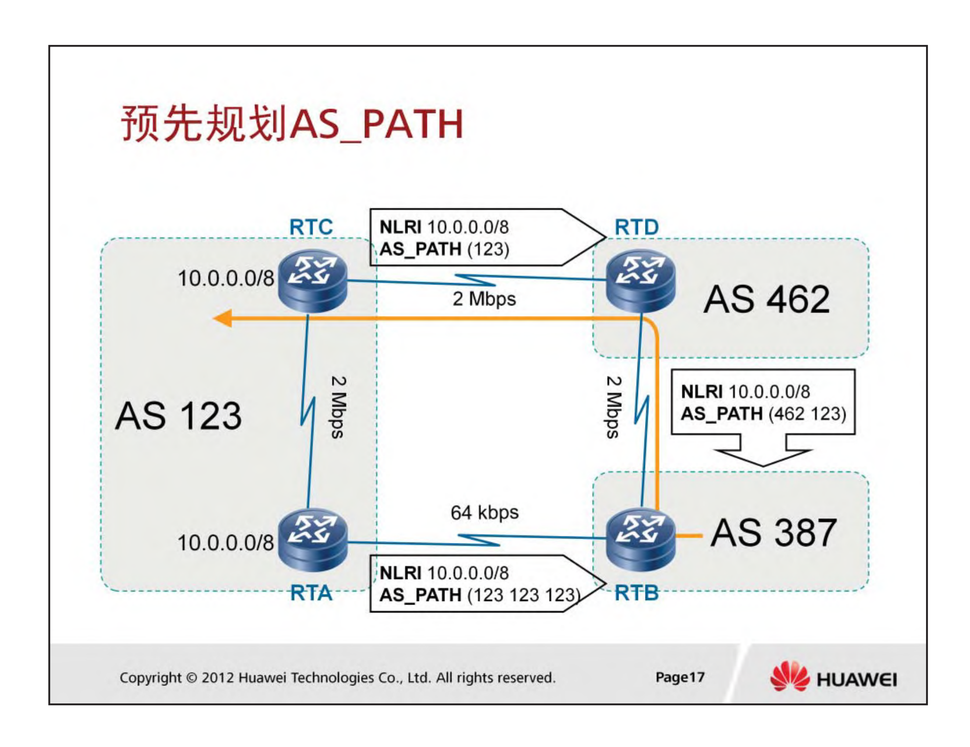
peer { group-name | ipv4-address } allow-as-loop [number]

参数：

group-name：对等体组的名称。

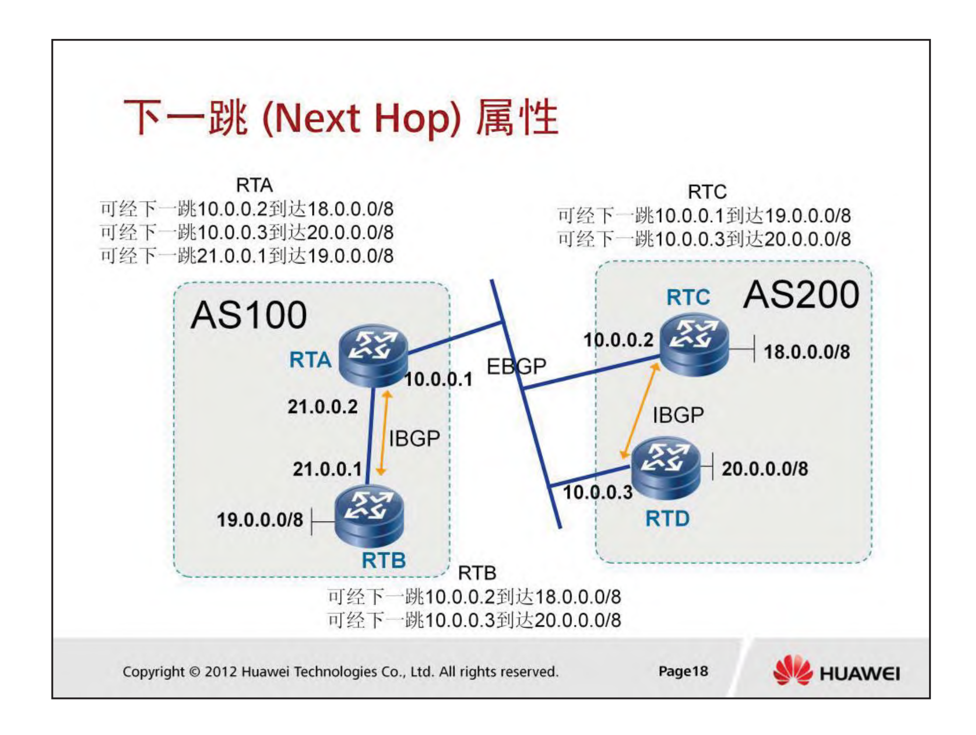
ipv4-address：对等体的IPv4地址。

number：本地AS号的的重复次数，范围为1～10。缺省值为1。



可以通过加长AS_PATH的列表长度，从而影响路径选择。

例如，从图中可以看到两条路径中RTC-RTD-RTB较RTA-RTB的带宽大，为避免选择带宽较小的路径，我们可以在RTA上配置，使路由10.0.0.0/8发往邻居时，将其AS_PATH属性再加上两个自治系统号123、123，这样当这条路由被传递到RTB时，其AS_PATH为：（123 123 123）。而从AS462传来的始发于AS123的路由的AS_PATH为（462 123）。这样RTB会比较AS_PATH的长度，最终路由器使用较优的路径：RTC-RTD-RTB。



下一跳属性是一个公认必遵属性。BGP中的下一跳概念稍微复杂，它可以是以下三种形式之一：

（注：图中的例子是，RTA与RTC通过直连以太网接口建立EBGP邻居关系，RTA与RTB通过直连接口建立IBGP邻居关系，而RTC与RTD通过直连以太网接口10.0.0.2和10.0.0.3建立IBGP邻居关系）

1、BGP在向EBGP邻居通告路由时，或者将本地发布的BGP路由通告给IBGP邻居时，下一跳属性是本地BGP与对端连接的端口地址。如胶片所示，RTC在向RTA通告路由18.0.0.0/8时，下一跳属性为10.0.0.2；RTB在向RTA通告路由19.0.0.0/8时，下一跳属性为21.0.0.1。

2、对于多路访问的网络（广播网或NBMA网络），下一跳情况有所不同：如胶片所示，RTC在向RTA通告路由20.0.0.0/8时，发现本地端口10.0.0.2同此路由的下一跳10.0.0.3（指在RTC路由表中此路由的下一跳）为同一子网，将使用10.0.0.3作为向EBGP通告路由的下一跳，而不是10.0.0.2。

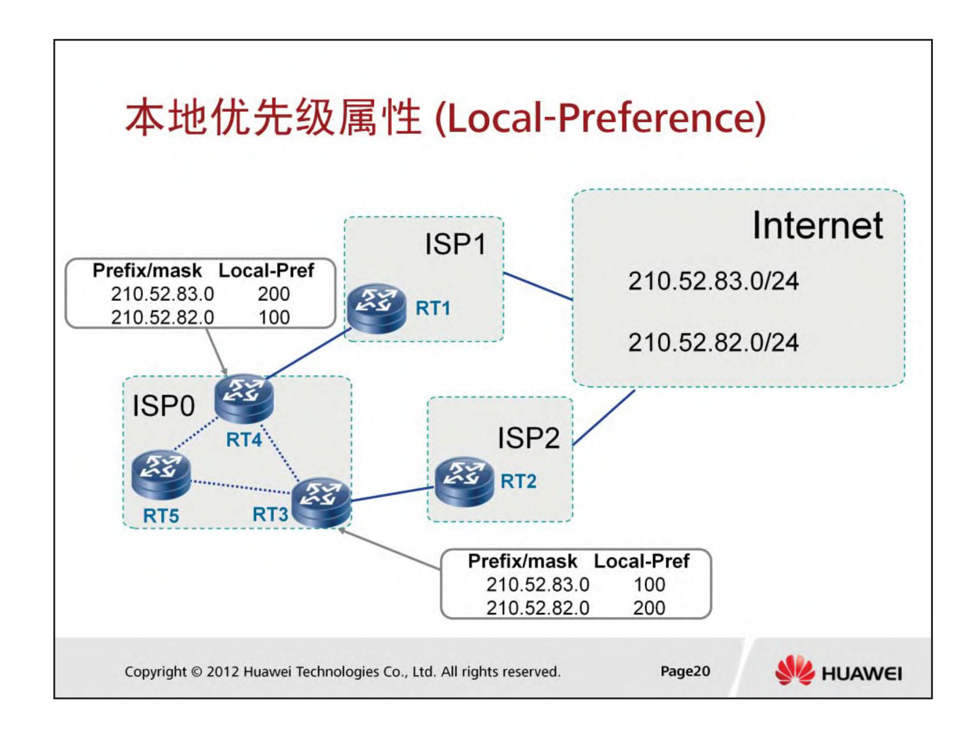
3、BGP在向IBGP通告从其它EBGP得到的路由时，不改变路由的下一跳属性，而直接传递给IBGP邻居。如胶片所示，RTA通过IBGP向RTB

通告路由18.0.0.0时，下一跳属性为10.0.0.2。这样做，有时会产生问题：如果RTB不知如何去往10.0.0.2，那么此BGP路由将失效。

解决方法：

方法一：可以在RTA的BGP视图下引入直连路由；

方法二：在RTA上，使用命令peer { group-name | ipv4-address } next-hop-local。此命令用来设置BGP向对等体组/对等体通告路由时，把下一跳属性设为自身的IP地址。



Local-Preference是公认任意属性。

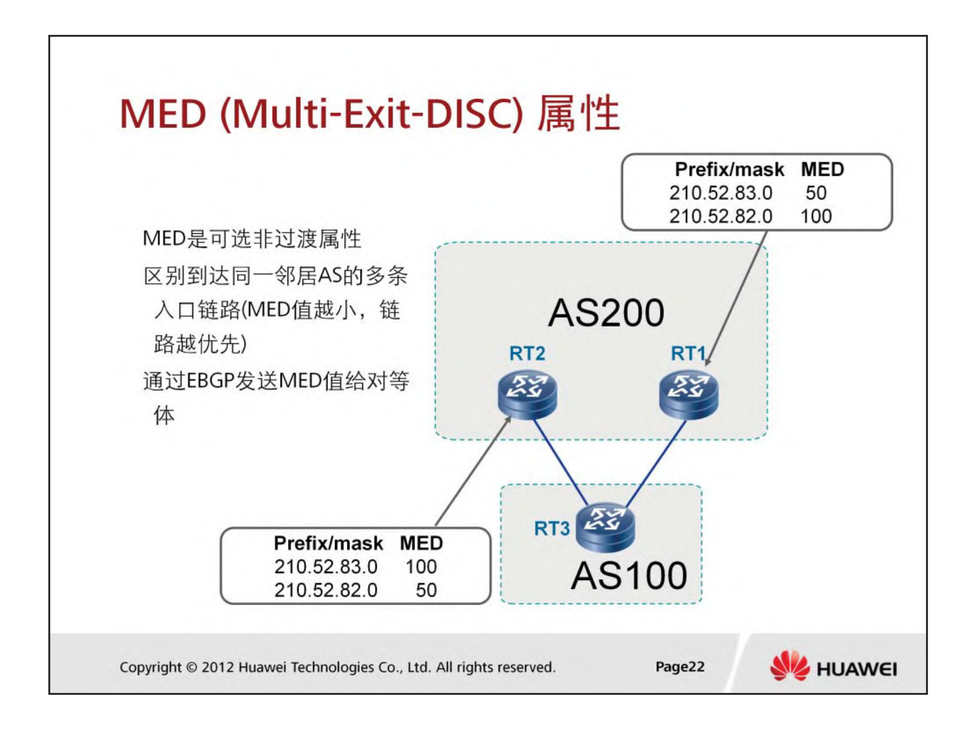
在某些情况下，一个ISP可能通过两条高速链路连接两个大的ISP作为自己到INTERNET的出口，如图所示，ISP0通过两条链路分别连接到ISP1和ISP2。

在这种情况下，ISP0怎样把流量均衡的分布到两条上行链路。假设INTERNET上有这样两条路由：210.52.83.0/24（在后面的介绍中以83代表）和210.52.82.0/24（在后面的介绍中以82代表），我们的目标是使到网络83的流量分布在到ISP1的链路上，而到网络82的流量分布在到ISP2的链路上。

分析ISP0内部网络结构，RT3，RT4和RT5之间分别两两建立TCP连接来构成IBGP对等体关系，而RT3和RT4分别和位于ISP2和ISP1的路由器建立EBGP对等体关系。这样路由器RT3和RT4都会从自己的EBGP对等体收到82和83这两条路由，而且RT3和RT4也会通过IBGP对等体关系通告82、83这两条路由给自己的IBGP对等体。由此可以看出，RT5分别有两个来源获得82和83路由，这样我们只需要在RT3和RT4上适当的对路由属性进行修改，就可以达到目的。

那么怎样做到这一点呢？在这里，BGP可以给路由附加一种称为本地优先级的属性，路由器接收到去往同一目的地的多条路由，可以判断本地优先级属性值的高低进行路由选择（本地优先级的数值越高越好）。

本例中，在RT3上，当从ISP2获得路由82和83的时候，给83赋予本地优先级属性100（默认，不需配置），而给82赋予本地优先级属性200；同样的道理，在RT4上，当从ISP1获得路由82和83的时候，给82赋予100而给83赋予200。这样对等体RT5就会从两个地方接收到了带有不同本地优先级属性值的同一目的地址的两条路由，根据本地优先级数值的高低进行路由选举。最终，实现到达83的流量分布在ISP1上，而到达82的流量分布在ISP2上。



前面介绍的本地优先级属性用于控制数据流怎样出AS，有些情况下，需要控制数据流怎样进入本AS，举一个例子。

在这个网络中，AS100通过两条上行链路连接AS200的两个不同的路由器，假设在AS200中有这样两个网络：210.52.83.0/24（在后面的介绍中以83代表）和210.52.82.0/24（在后面的介绍中以82代表），这两个网络都通过BGP协议通告给了AS100的边界路由器RT3。这时候，AS200的管理者想达到这样一个目的：从AS100来的到82的数据流通过RT2路由器到达，而从AS100来的到83的数据流通过RT1到达。可以看出，跟前面在AS内部控制数据流的出口不同的是，我们需要在AS内部控制数据流怎样流入该AS。

跟前面的思路相同，我们还是给通告的路由一种标记，当对端接收到多条去往同一网段的路由时，根据该标记决定选择哪条路由。

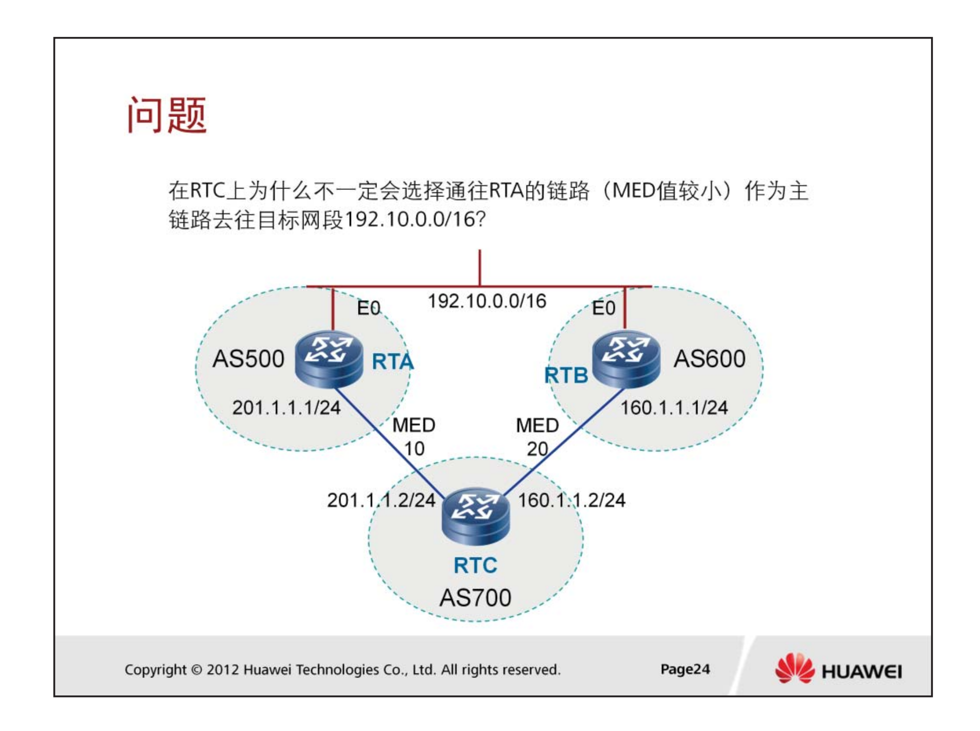
在AS200的边界路由器RT1上，当向RT3发布路由82和83时，给83打上标记50，而给82打上标记100；

在AS200的边界路由器RT2上，当向RT3发布路由82和83时，给82打上标记50，而给83打上标记100；

当AS100路由器RT3通过EBGP对等体分别从RT1和RT2获得去往相

同网段的路由时，会选择RT1作为83的下一跳而选择RT2作为82的下一跳。

这种标记我们也用属性的方式实现，这个标记是一个整数，数值越小，在选择中越有优势，我们称这种标记为MED（外部度量）。可以看出，跟本地优先级不同的是，MED控制流量怎样进入AS，而本地优先级则控制流量怎样流出AS。



缺省情况下，不允许比较来自不同AS邻居的路由信息的MED值。但是，我们可以通过配置compare-different-as-med命令来允许比较来自不同自治系统中的邻居的路由的MED值。不过，除非能够确认不同的自治系统采用了同样的IGP和路由选择方式，否则不要使用此命令。

团体 (Community) 属性

什么是团体属性

- 团体是一组有相同性质的目的地址路由。目的就是将路由信息编组，通过组的标识决定路由传递的策略。

团体属性

- 属性类型：8
- 可选，过渡属性

 Page25



在BGP的范围内，一个团体是一组有公共性质的目的地址。RFC1997定义了团体路径属性是一个可变长度的可选，过渡属性。

每个AS的管理员都可以自己定义目的地址所属的团体，默认情况下，所有目的路由都属于常规Internet团体。

一条路由可以具有一个以上的团体属性值。如果在一条路由中包含有多个团体属性值，BGP路由器可以根据一个、一些或所有这些属性值来采取相应的策略。路由器在将路由传递给其他对等体之前可以增加或修改团体属性值。

团体 (Community) 属性

团体属性是由一系列4字节(0x00000000—0xFFFFFFFF)数值所组成

- 保留的团体属性：
  - 0x00000000—0x0000FFFF
  - 0xFFFF0000—0xFFFFFFFF
- 公认团体属性：
  - NO_EXPORT (0xFFFFFF01)
  - NO_ADVERTISE (0xFFFFFF02)
  - NO_EXPORT_SUBCONFED (0xFFFFFF03)
- 私有团体属性：
  - AS(2B):Number(2B)

 Page26



团体属性由一系列以4字节长度为单位的数值所组成，每4字节代表一个团体属性。所有路由的团体属性都属于团体属性列表。

团体属性数值定义从0x00000000到0x0000FFFF和从0xFFFF0000到0xFFFFFFFF被保留。

公认团体属性是公认的，具有全球意义。公认的团体有：

NO_EXPORT(0xFFFFFF01)：路由器收到带有这一团体值的路由后，不应把该路由通告给一个联盟之外的对等体。

NO_ADVERTISE(0xFFFFFF02)：路由器收到带有这一团体值的路由后，不应把该路由通告给任何的BGP对等体。

NO_EXPORT_SUBCONFED(0xFFFFFF03)：路由器收到带有这一团体值的路由后，可以把该路由通告给它的IBGP对等体，但不应通告给任何的EBGP对等体（包括联盟内的EBGP对等体）。

除了这些公认的团体属性值外，私有的团体属性值也可以被定义用于特殊用途。这些属性值被一些数字所标示。通常都是前2字节由本地AS来编码，后2字节是一个0到65535之间的任意数值。(例如：AS690被定义为研发、教育和商务部所使用，团体属性数值应该被定义在0x02B20000到0x02B2FFFF(690:0～65535）之间)


目 录
BGP路径属性介绍
BGP路径选择
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page28
HUAWEI

BGP路径选择过程

1，如果此路由的下一跳不可达，忽略此路由

2，Preferred-Value值数值高的优先

3，Local-Preference值最高的路由优先

4，聚合路由优先于非聚合路由

5，本地手动聚合路由的优先级高于本地自动聚合的路由

6，本地通过**network**命令引入的路由的优先级高于本地通过**import-route**命令引入的路由

7，AS路径的长度最短的路径优先

8，比较Origin属性，IGP优于EGP，EGP优于Incomplete

9，选择MED较小的路由

 Page29



1.如果此路由的下一跳不可达，忽略此路由

2.Preferred-Value值数值高的优先（VRP5增加的新参数，指定对等体的首选值，数值越高越好）

3.Local-Preference值最高的路由优先

4.聚合路由优先于非聚合路由

5.本地手动聚合路由的优先级高于本地自动聚合的路由

6.本地通过network命令引入的路由的优先级高于本地通过import-route命令引入的路由

7.AS路径的长度最短的路径优先

8.比较Origin属性， IGP优于EGP，EGP优于Incomplete

9.选择MED较小的路由

BGP路径选择过程 (续)

10，EBGP路由优于IBGP路由

11，BGP优先选择到BGP下一跳的IGP度量最低的路径

- 当以上全部相同，则为等价路由，可以负载分担
  - 注：AS_PATH必须一致
  - 当负载分担时，以下3条原则无效

12，比较Cluster-List长度，短者优先,

13，比较Originator_ID(如果没有Originator_ID，则用Router ID比较)，选择数值较小的路径

14，比较对等体的IP地址，选择IP地址数值最小的路径

 Page30



10. EBGP路由优于IBGP路由
11. EBGP优先选择到BGP下一跳的IGP度量最低的路径
    - 当以上全部相同，则为等价路由，可以负载分担
      - 注：AS_PATH必须完全一致
      - 当负载分担时，以下3条原则无效
12. 比较Cluster-List长度，短者优先
13. 比较Originator_ID(如果没有Originator_ID，则用Router ID比较)，选择数值较小的路径
14. 比较对等体的IP地址，选择IP地址数值最小的路径



问 题

BGP路径属性的作用？BGP发展到现在为止总共有多少种属性?

AS_PATH属性怎样防止路由环路?

MED与LOCAL_PREF属性的区别是什么?

 Page31



1.BGP路径属性的作用？BGP发展到现在为止总共有多少种属性?

答：BGP是一个路由选择策略工具，区别于IGP，BGP目的是传递路由而不是发现路由。而BGP的路径属性就是一组描述BGP前缀特性的参数，让BGP可以很好的控制路由信息的传递。而BGPv4发展到现在为止，总共有16种路径属性。

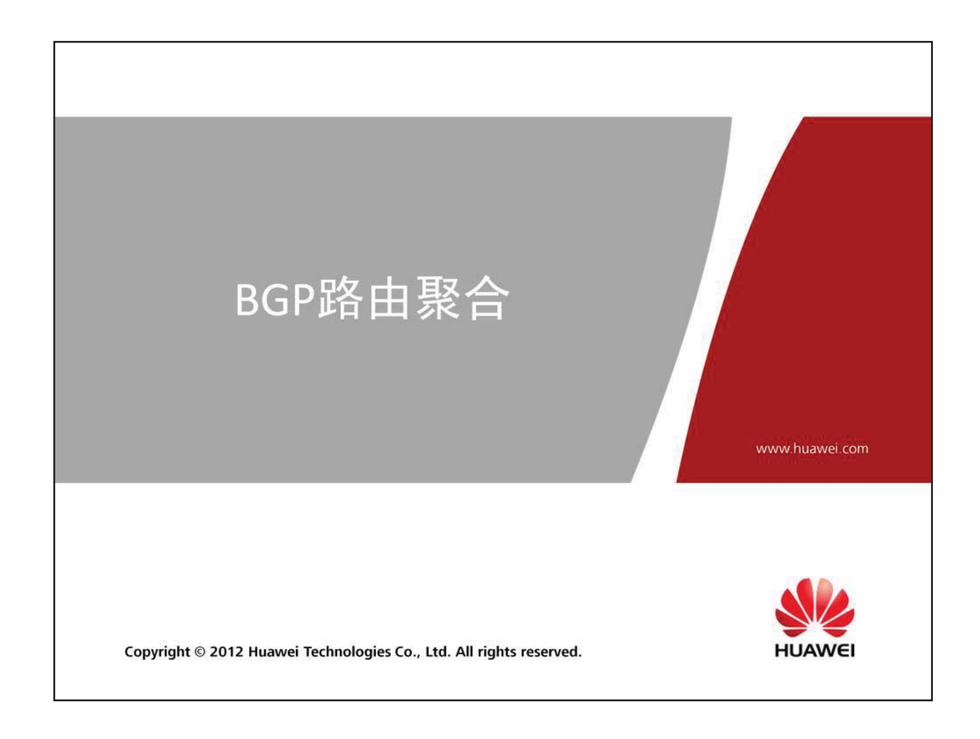
2.AS_PATH属性是怎样防止路由环路?

答：AS_PATH属性列表以相反的顺序列出了一条前缀先后经过的AS，每经过的一个AS，该AS号会被放入AS_PATH列表的最前面(开始处)。如果在路由更新消息中发现自己所在的AS号已经被包含在AS_PATH属性中，则表明该路由之前曾经通过该AS或者是源自于该AS，为避免路由环路，将此路由信息丢弃。

3.MED与LOCAL_PREF属性的区别是什么?

答：MED主要作用在EBGP对等体上，而LOCAL_PREF则主要作用在IBGP对等体上。换种说法就是，MED通常通过EBGP对等体向外发送，从而实现对对端AS入流量的控制；而LOCAP_PREF通常向IBGP对等体发送，从而实现对本地AS出流量的控制。


谢谢
www.huawei.com


BGP路由聚合
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

BGP作为一个跨域路由协议，在很多情况下都需要对明细路由进行聚合，并通告到其它远端的AS，而我们将会在本课程中讨论BGP聚合后的路径属性变化情况以及聚合后所引起的问题。

Page1





培训目标

学完本课程后，您应该能：

- 知道BGP路由聚合的方法
- 了解AS_PATH属性的变化
- 知道BGP路由聚合配置
- 知道BGP路由聚合策略

Page2



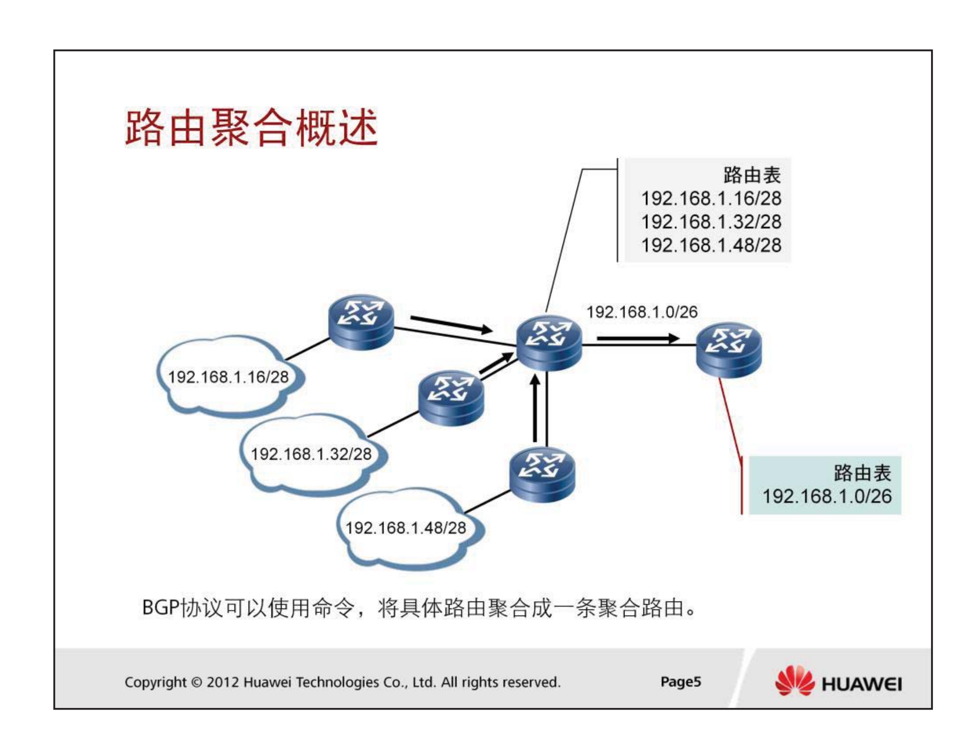


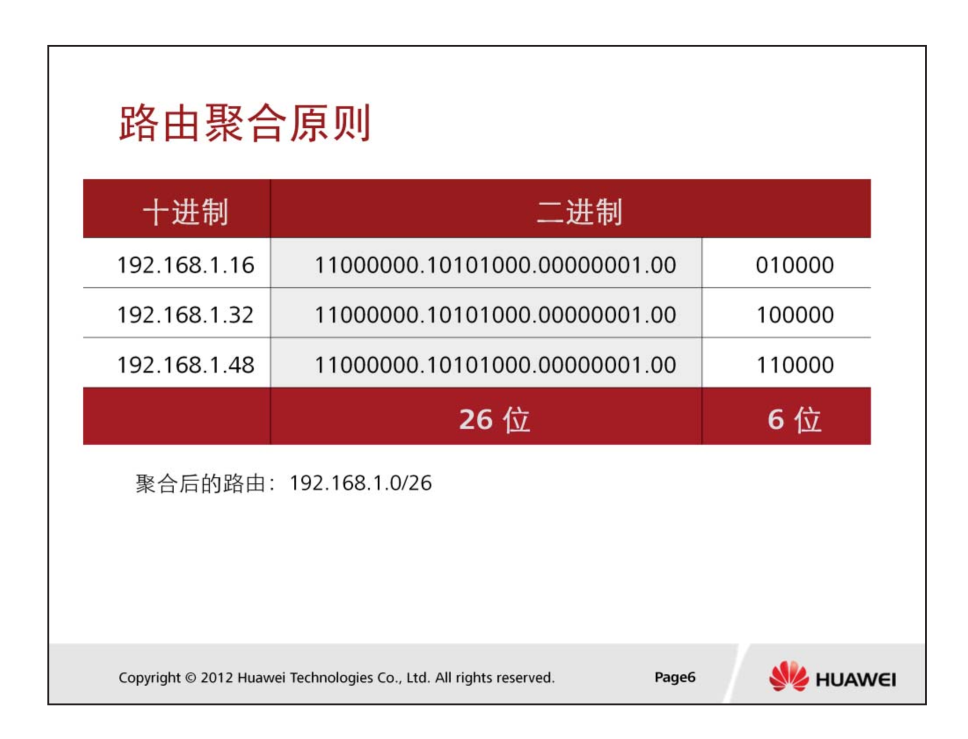
什么是路由聚合

AS_SET

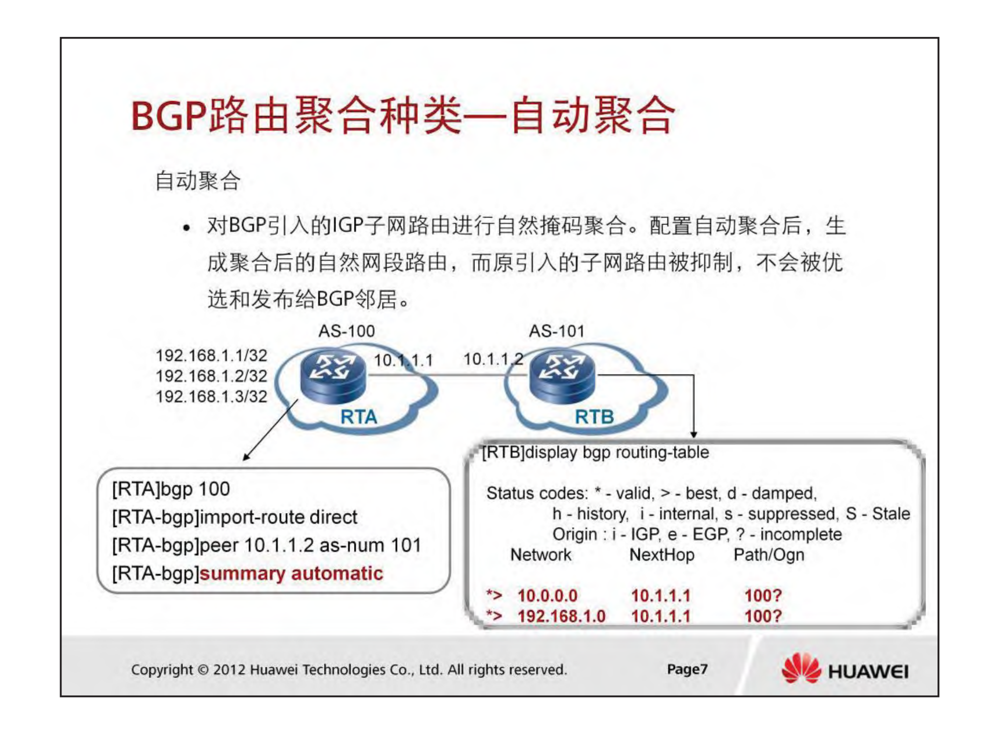
改变路由聚合属性




路由聚合概述
路由表
192.168.1.16/28
192.168.1.32/28
192.168.1.48/28
192.168.1.0/26
192.168.1.16/28
192.168.1.32/28
192.168.1.48/28
路由表
192.168.1.0/26
BGP协议可以使用命令，将具体路由聚合成一条聚合路由。
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page5
HUAWEI



路由聚合原则采用最长相同掩码匹配的方法。

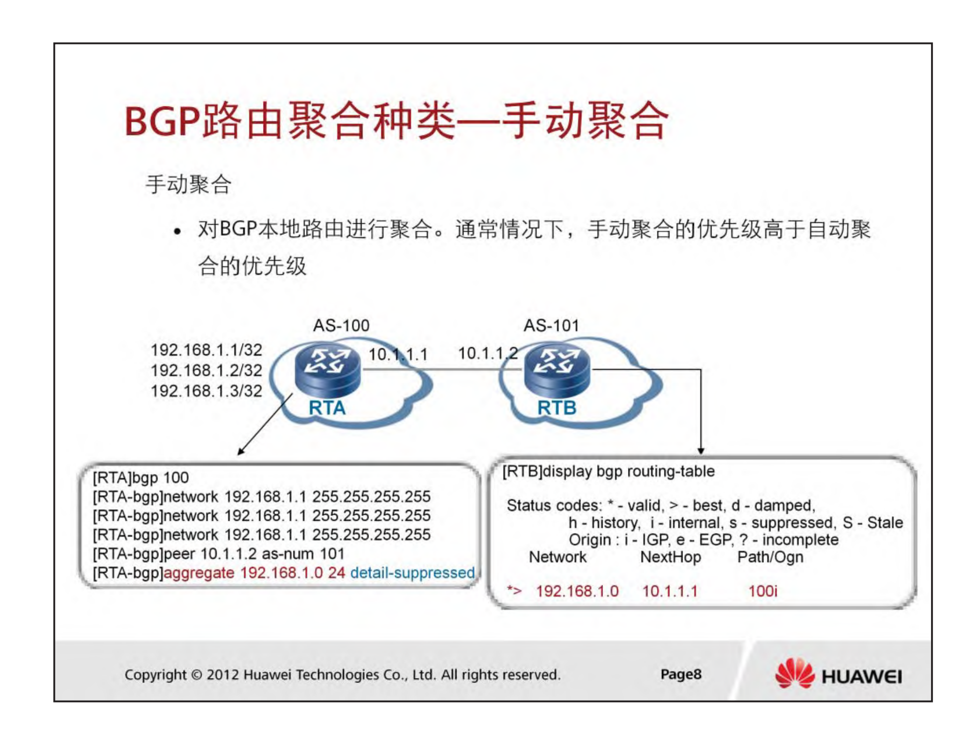


自动聚合:

对BGP引入的IGP子网路由进行聚合。配置自动聚合后，生成聚合后的自然网段路由，而原引入的子网路由被抑制，不会被优选和发布给BGP邻居。

summary automatic命令用来使能对引入的路由进行自动聚合，undo summary automatic命令用来取消对引入的路由进行自动聚合。

缺省情况下，不对引入的路由进行自动聚合。



手动聚合：

对BGP本地路由进行聚合。通常情况下，手动聚合的优先级高于自动聚合的优先级。缺省情况下手动聚合后会把明细路由和聚合路由一起发布。

aggregate命令用来在BGP路由表中创建一条聚合路由，undo aggregate命令用来关闭该功能。

缺省情况下，不进行路由聚合。

通过“aggregate”命令把多条BGP明细路由聚合为一条汇总路由，并通告给其它对等体。与summary automatic命令不一样，aggregate命令需要手动输入指定聚合的前缀和掩码 ip-address mask [as-set | attribute-policy route-policy-name1 | detail-suppressed | origin-policy route-policy-name2 | suppress-policy route-policy-name3]

BGP路由聚合需要考虑的问题

BGP路由聚合需要考虑的问题

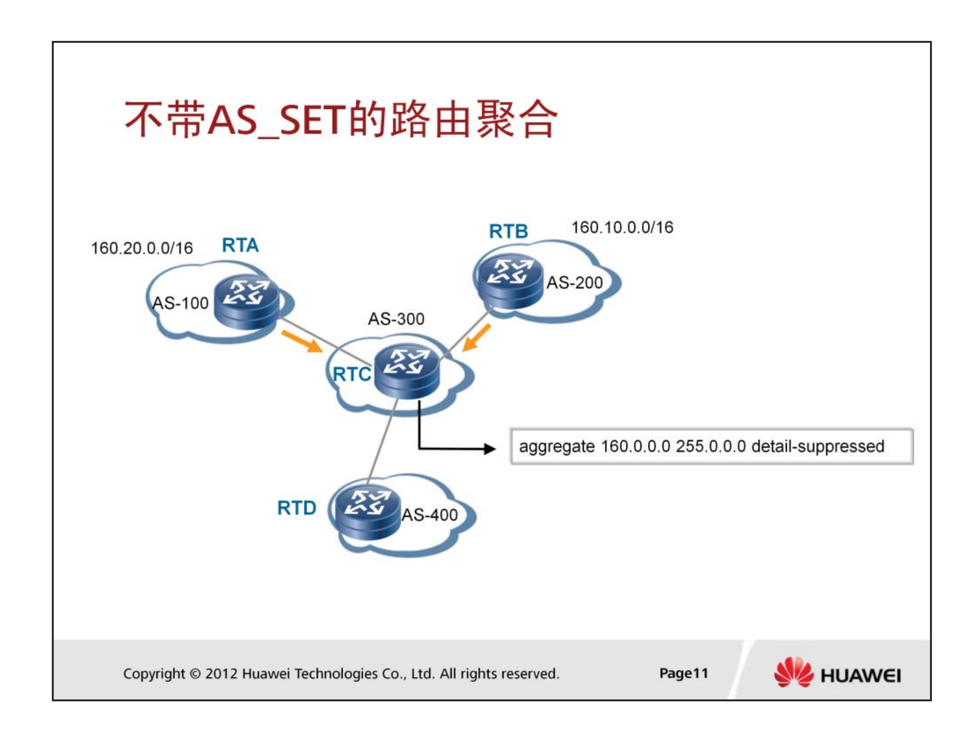
- 明细路由的发布
- BGP路由属性的继承
  - AS-Path
  - Origin
  - Community
  - … …





Page10





AS_Path属性按一定次序记录了某条路由从本地到目的地址所要经过的所有AS编号。

华为产品VRP支持4种类型的AS_PATH属性:

AS_SEQUENCE

AS_SET

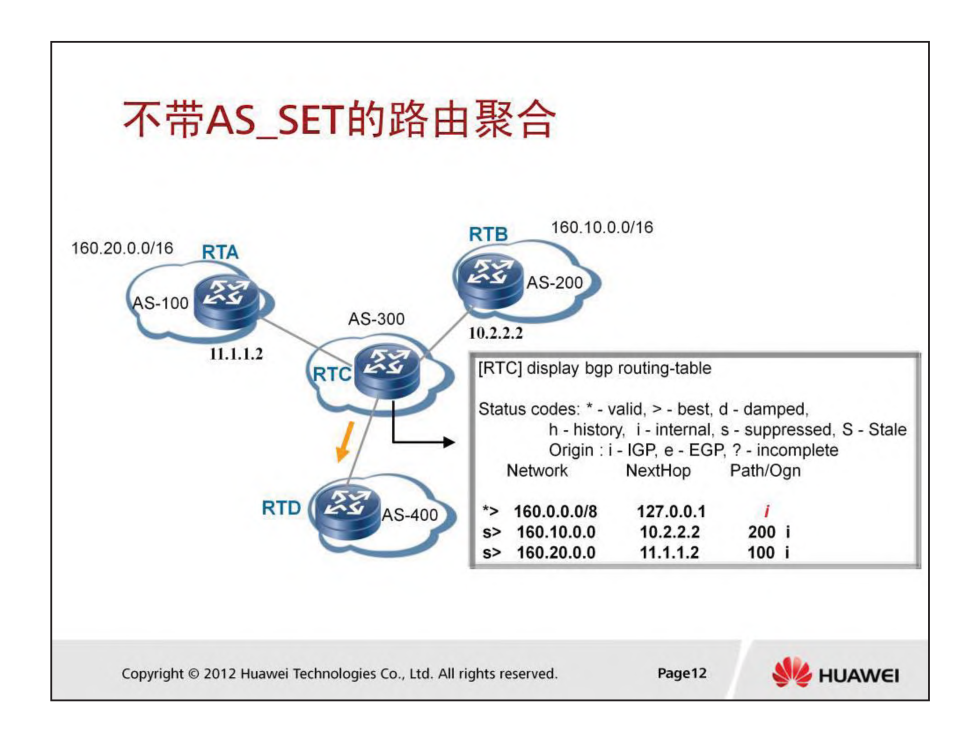
AS_CONFED_SEQUENCE

AS_CONFED_SET

SET和SEQUENCE的不同之处在于，SET选项下的AS列表通常用于路由聚合，将来自不同AS的AS号无序排列在AS列表里；而SEQUENCE选项下的AS列表是有序的，每经过一个AS都会将其AS号排列在列表的前端。

AS_CONFED仅仅只能应用于BGP联盟的情况下，一旦路由信息向外部AS更新时，AS_CONFED将会被删去。

当RTA与RTB把本地网络宣告给RTC，RTC通过命令：aggregate 160.0.0.0 255.0.0.0 detail-suppressed 聚合为160.0.0.0/8，然后通告给RTD。



通过display bgp routing-table查看RTC的路由表：

Status codes: * - valid, > - best, d - damped,

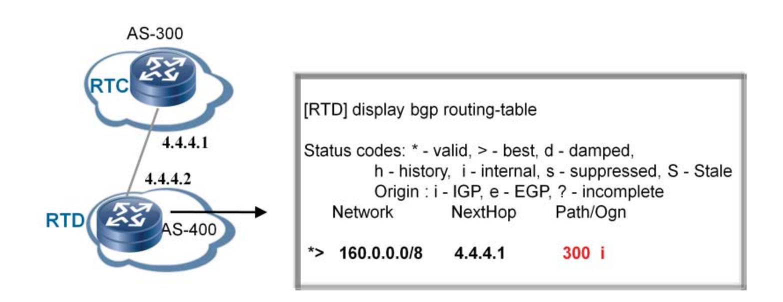
h - history, i - internal, s - suppressed, S - Stale

Origin : i - IGP, e - EGP, ? - incomplete

	Network	NextHop	Path/Ogn
*>	160.0.0.0/8	127.0.0.1	i
s>	160.10.0.0	10.2.2.2	200 i
s>	160.20.0.0	11.1.1.2	100 i

可以看到，通过命令聚合后的BGP路由表多了一条聚合路由，并且该聚合路由的AS-Path属性里没有任何其它AS信息，说明没有带AS_SET参数的聚合路由会被认为是由RTC产生的。





聚合路由 160.0.0.0/8 被认为是始发于 AS-300，并且丢失了所有具体路由 160.10.0.0/16 和 160.20.0.0/16 的AS-PATH信息。

 Page13



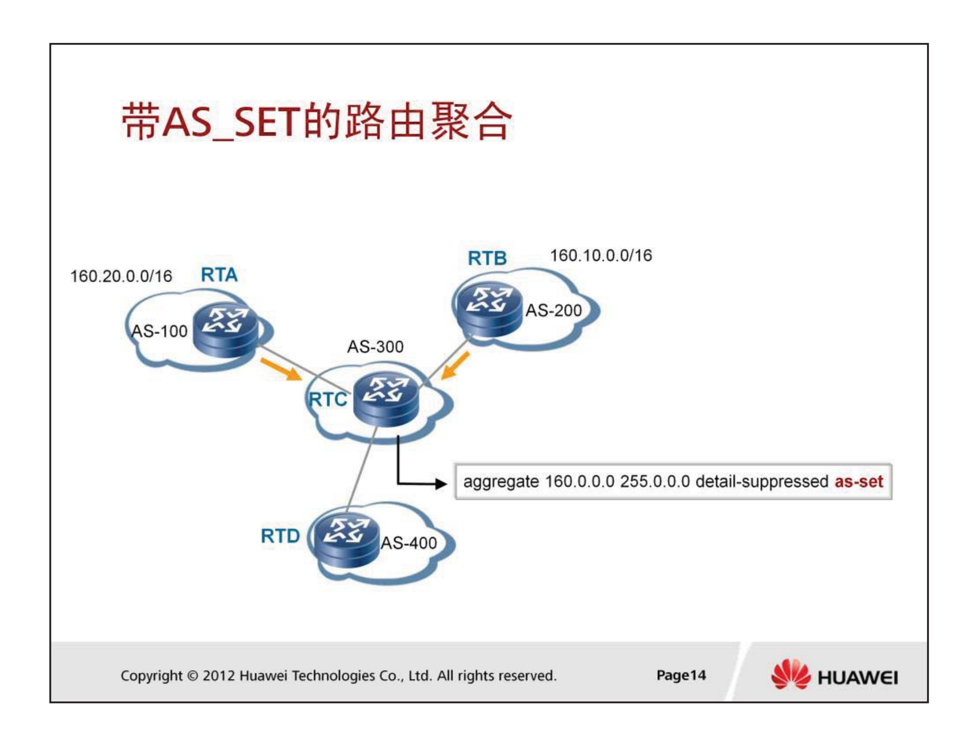
RTC将来自于 AS_100的路由160.20.0.0/16和来自于AS_200的路由160.10.0.0/16 进行聚合，因为 RTC上配置了参数detail-suppressed，这样只有聚合路由被发布到 RTD。具体路由160.10.0.0/16 和160.20.0.0/16 被抑制，下面是 RTD上BGP路由表的信息，注意聚合路由的as-path属性。

```
[RTD] display bgp routing-table

Status codes: * - valid, > - best, d - damped,
          h - history,  i - internal, s - suppressed, S - Stale
          Origin : i - IGP, e - EGP, ? - incomplete
    Network         NextHop      Path/Ogn

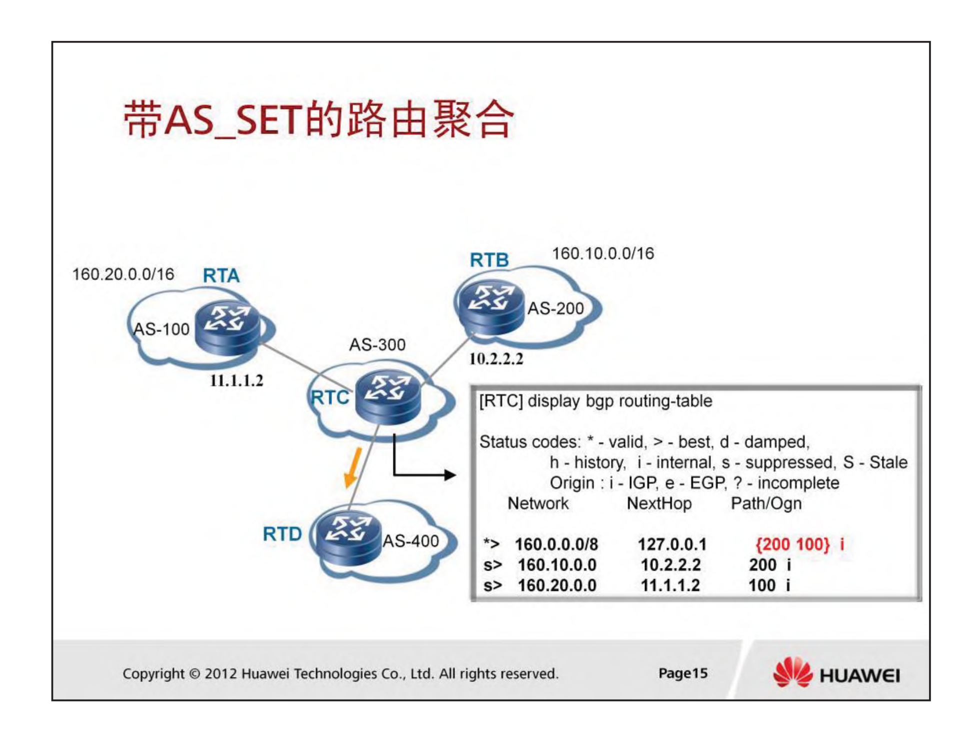
 *>  160.0.0.0/8      4.4.4.1       300  i
```

聚合路由 160.0.0.0/8 被认为是始发AS-300，Origin属性为 IGP，并且丢失了所有具体路160.10.0.0/16 和 160.20.0.0/16 的as-path 信息。



现在我们在 RTC 的聚合命令中增加 as-set 参数。配置如下：

aggregate 160.0.0.0 255.0.0.0 detail-suppressed as-set



改变配置后 RTC 上的 BGP 路由表如下：

[RTC] display bgp routing-table

Status codes: * - valid, > - best, d - damped,

h - history, i - internal, s - suppressed, S - Stale

Origin : i - IGP, e - EGP, ? - incomplete

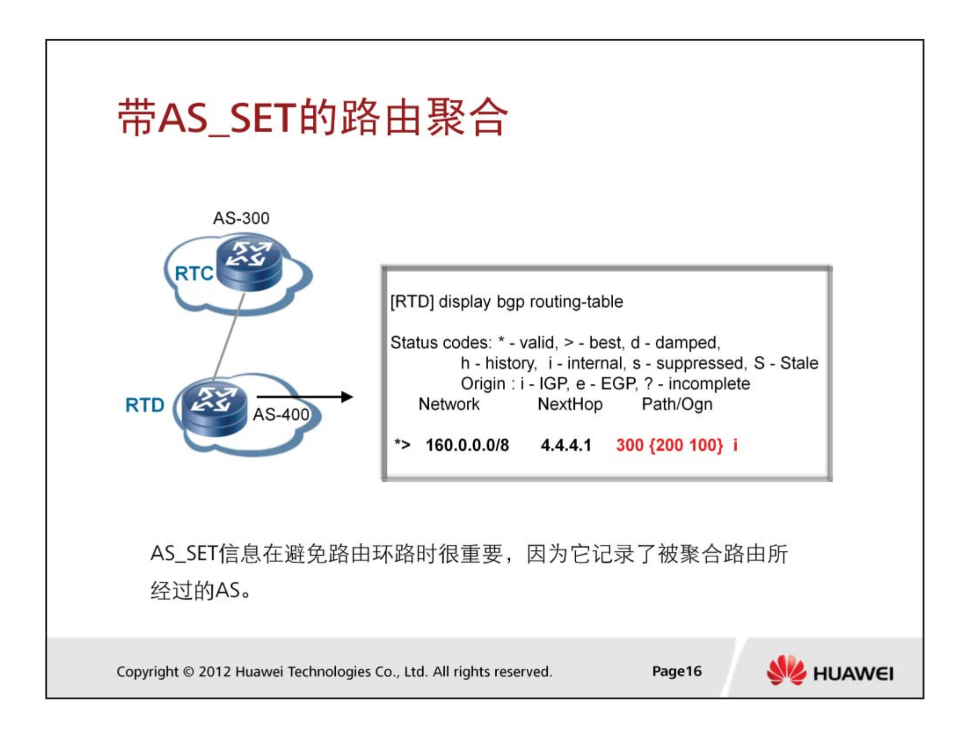
Network NextHop Path/Ogn

*> 160.0.0.0/8 127.0.0.1 {200 100} i

s> 160.10.0.0 10.2.2.2 200 i

s> 160.20.0.0 11.1.1.2 100 I

使用 as-set 参数后，RTC上 BGP 路由表中聚合路由的路径信息变为 {200，100}。它表明聚合操作聚合了来自于 AS-200 和 AS-100的路由。



改变配置后 RTD 上的 BGP 路由表如下:

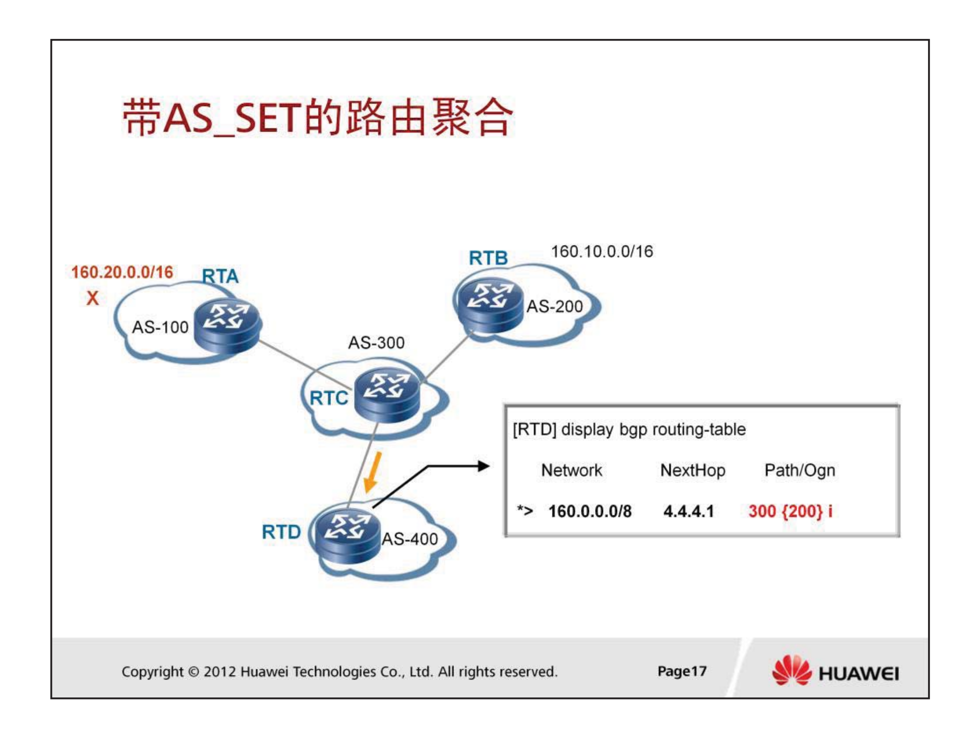
```
[RTD] display bgp routing-table

Status codes: * - valid, > - best, d - damped,
              h - history,  i - internal, s - suppressed, S - Stale
              Origin : i - IGP, e - EGP, ? - incomplete
     Network          NextHop        Path/Ogn

 *>  160.0.0.0/8      4.4.4.1        300 {200 100}  i
```

AS_SET 信息在避免路由环路时很重要，因为它记录了被聚合路由所经过的AS。

在闭合网络中，聚合路由可能通过 BGP 路由重新进入 AS_SET 中列出的任何一个 AS，这样就可能形成环路，BGP的环路检测机制检测到自己的AS 号在聚合路由的AS_SET 属性列出的 AS中，就会丢弃该聚合路由，这样就避免了形成环路。



使用as-set 参数后聚合路由的 AS 信息中包含被聚合的每条具体路由的 AS 信息，并随着被聚合路由的更新而变化。在上面的例子中，如果路由160.20.0.0/16 被撤销，聚合路由的路径信息将从300{200，100} 变为300 {200}，聚合路由的属性发生了变化。如果聚合路由聚合了成千上万条路由，而且具体路由有问题的话，聚合路由就会不断地发生振荡。


目 录
什么是路由聚合
AS_SET
改变路由聚合属性
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page18
HUAWEI

聚合路由的问题

聚合路由不继承原有BGP路由属性

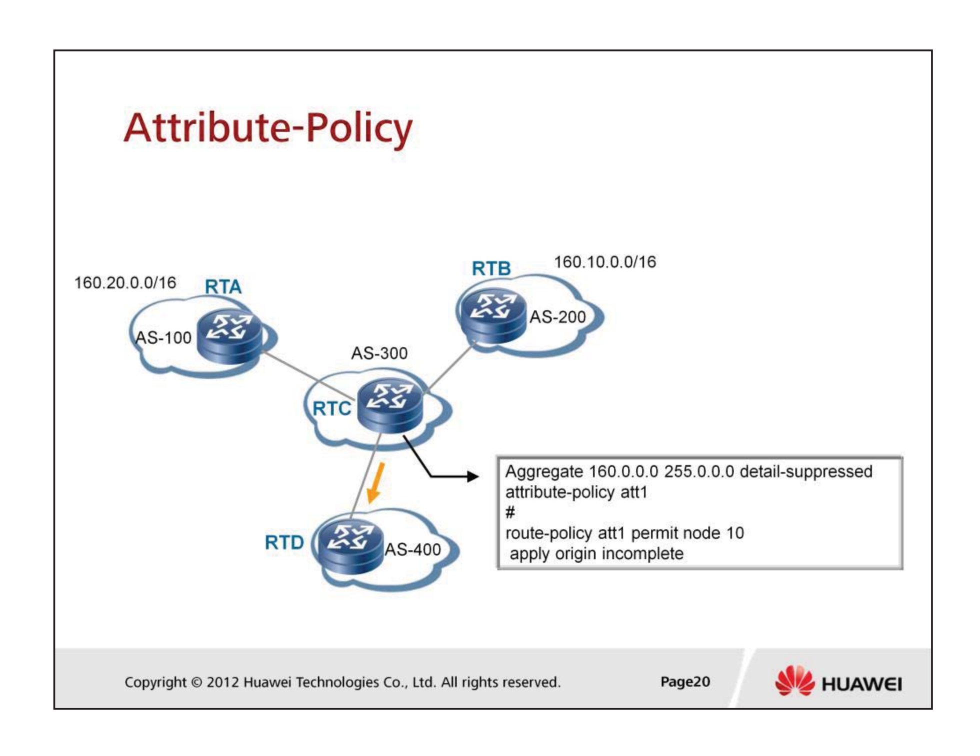
可以通过命令修改聚合路由属性：

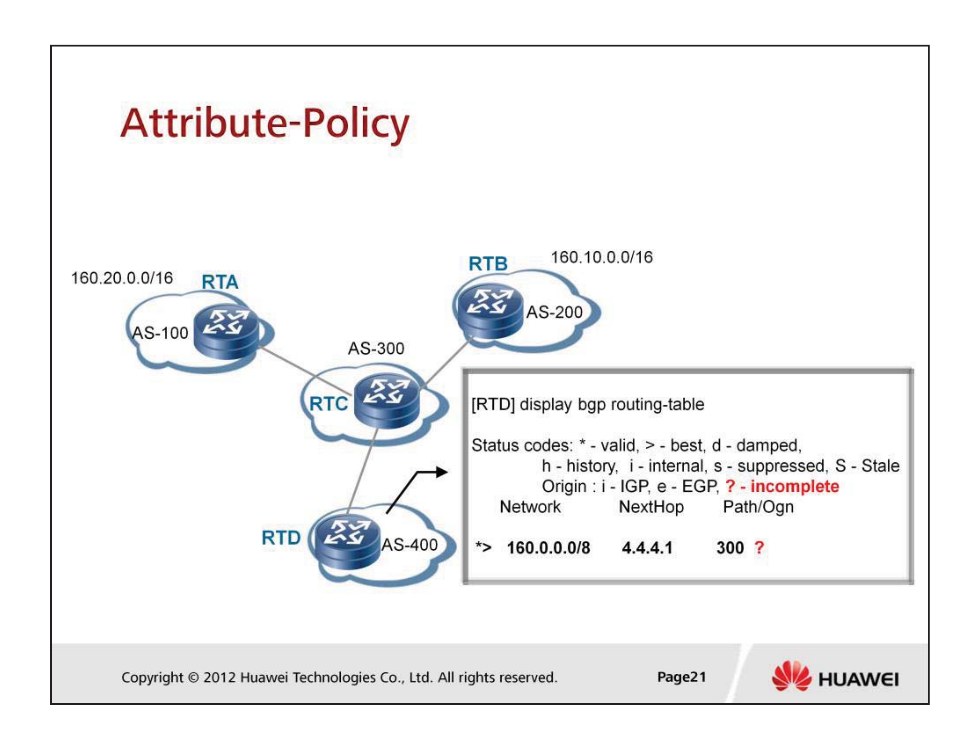
aggregate ***ip-address mask*** [as-set | attribute-policy ***route-policy-name1*** | detail-suppressed | origin-policy ***route-policy-name2*** | suppress-policy ***route-policy-name3***]



BGP聚合路由默认的情况下，不继承原有BGP路由的属性。

通过命令aggregate ip-address mask [as-set | attribute-policy route-policy-name1 | detail-suppressed | origin-policy route-policy-name2 | suppress-policy route-policy-name3]， 配置attribute-policy参数可用于修改BGP聚合路由属性。


Attribute-Policy
160.20.0.0/16
RTA
AS-100
RTB
160.10.0.0/16
AS-200
AS-300
RTC
RTD
AS-400
Aggregate 160.0.0.0 255.0.0.0 detail-suppressed
attribute-policy att1
#
route-policy att1 permit node 10
apply origin incomplete
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page20
HUAWEI


Attribute-Policy
160.20.0.0/16
RTA
AS-100
RTB
160.10.0.0/16
AS-200
AS-300
RTC
RTD
AS-400
[RTD] display bgp routing-table
Status codes: * - valid, > - best, d - damped,
h - history, i - internal, s - suppressed, S - Stale
Origin : i - IGP, e - EGP, ? - incomplete
Network NextHop Path/Ogn
*> 160.0.0.0/8 4.4.4.1 300 ?
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page21
HUAWEI

过滤策略

origin-policy

- 使用关键字**origin-policy**仅选择符合route-policy的具体路由来生成聚合路由。

suppress-policy

- 关键字**suppress-policy**能产生聚合路由，但抑制指定路由的通告。可以用**route-policy**的**if match**子句有选择地抑制一些具体路由，其它具体路由仍被通告。

Page22


origin-policy

- 使用关键字origin-policy仅选择符合route-policy的具体路由来生成聚合路由。

suppress-policy

- 关键字suppress-policy能产生聚合路由，但抑制指定路由的通告。可以用route-policy的if match子句有选择地抑制一些具体路由，其它具体路由仍被通告。





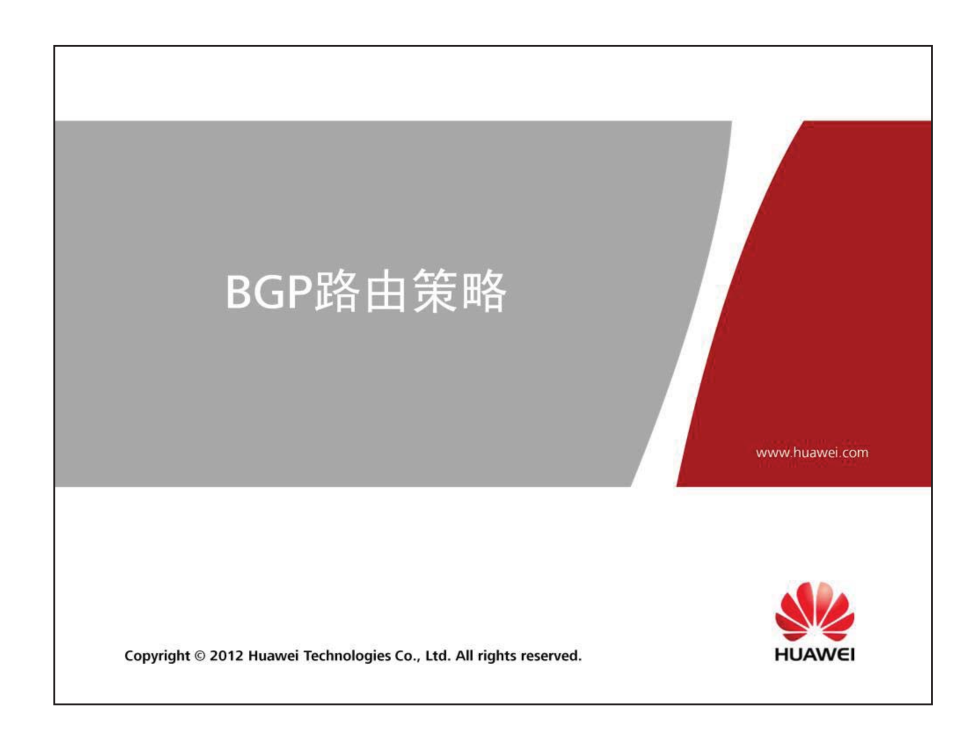
1.BGP聚合路由中，AS_SET的作用?

答：生成具有AS-SET的聚合路由。

2.请说出BGP聚合时需要注意的事项?

答：需要注意聚合路由的属性改变，比如说AS_PATH属性，聚合以后如果没加上AS_SET参数，系统会认为该聚合路由是由聚合路由器所产生，并以本AS号代替所有其它的AS号；除AS_PATH属性需要注意以外，其它BGP的属性都需要注意。


谢谢
www.huawei.com


BGP路由策略
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

BGP可以结合几乎所有的策略工具，并利用BGP路径属性，如：AS_PATH， COMMUNITY等，过滤从邻居收到或发送给邻居的路由信息。通过本课程的学习，您最终会发现，BGP由始至终都是一个策略工具。

Page1





培训目标

学完本课程后，您应该能：

- 了解BGP策略选路
- 知道BGP过滤应用

Page2




目 录
BGP选路原则回顾
实现BGP路径选择
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page4
HUAWEI

BGP选路原则回顾

1.如果此路由的下一跳不可达，忽略此路由

2.Preferred-Value值数值高的优先

3.Local-Preference值最高的路由优先

4.聚合路由优先于非聚合路由

5.本地手动聚合路由的优先级高于本地自动聚合的路由

6.本地通过**network**命令引入的路由的优先级高于本地通过**import-route**命令引入的路由

7.AS路径的长度最短的路径优先

8.比较Origin属性，IGP优于EGP，EGP优于Incomplete

9.选择MED较小的路由

 Page5



1.如果此路由的下一跳不可达，忽略此路由

2.Preferred-Value值数值高的优先(VRP5增加的新参数，指定对等体的首选值，数值越高越好)

3.Local-Preference值最高的路由优先

4.聚合路由优先于非聚合路由

5.本地手动聚合路由的优先级高于本地自动聚合的路由

6.本地通过network命令引入的路由的优先级高于本地通过import-route命令引入的路由

7.AS路径的长度最短的路径优先

8.比较Origin属性， IGP优于EGP，EGP优于Incomplete

9.选择MED较小的路由

BGP选路原则回顾 (续)

10.BGP优先选择到BGP下一跳的IGP度量最低的路径

- 当以上全部相同，则为等价路由，可以负载分担
  - 注：AS-Path必须一致
  - 当负载分担时，以下3条原则无效

11.比较Cluster List长度，短者优先

12.比较Originator_ID(如果没有Originator_ID，则用Router ID比较)，选择数值较小的路径

13.比较对等体的IP地址，选择IP地址数值最小的路径

 Page6



10. EBGP路由优于IBGP路由
11. EBGP优先选择到BGP下一跳的IGP度量最低的路径
    - 当以上全部相同，则为等价路由，可以负载分担
      - 注：AS_PATH必须完全一致
      - 当负载分担时，以下3条原则无效
12. 比较Cluster-List长度，短者优先
13. 比较Originator_ID(如果没有Originator_ID，则用Router ID比较)，选择数值较小的路径
14. 比较对等体的IP地址，选择IP地址数值最小的路径


目 录
BGP选路原则回顾
实现BGP路径选择
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page7
HUAWEI

BGP选路参数

影响BGP选路的重要参数

- Preferred Value
- **Local-Preference**
- **AS-Path**
- Origin
- **MED**
- EBGP/IBGP
- IGP Cost
- Cluster List
- **Communities**

 Page8



根据BGP的选路原则以及BGP常用的路径属性，我们可以总结出9个影响BGP选路的重要参数，分别为：

Preferred Value

LOCAL_PREF

AS_PATH

ORIGIN

MED

邻居类别是EBGP还是IBGP

IGP内部开销值

Cluster List /ROUTER_ID

COMMUNITY

以上参数都能直接地影响BGP的路径选择，其中我们常用的参数分别为LOCAL_PREF, AS_PATH和MED属性。对于这些参数的配置以及使用方法，我们会在后面的课程里给大家详细介绍。

BGP Local-Preference

default local-preference命令用来配置BGP的缺省本地优先级，该值越大则优先级越高。

[Router-bgp] **default local-preference** ***preference***

缺省情况下，BGP本地优先级的值为100。

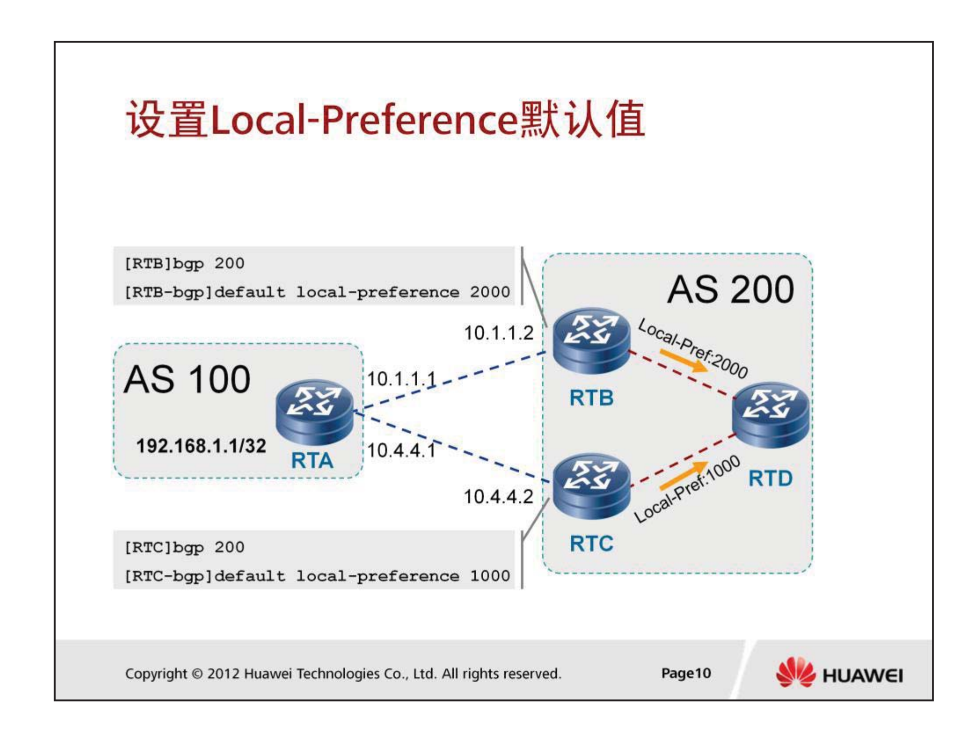
配置不同本地优先级会影响BGP的路由选择。当一个运行BGP的路由器有多条路由到达同一目的地址时，会优先选择本地优先级最高的路由。

本地优先级属性仅在IBGP对等体之间交换，不通告给其他AS。

 Page9 

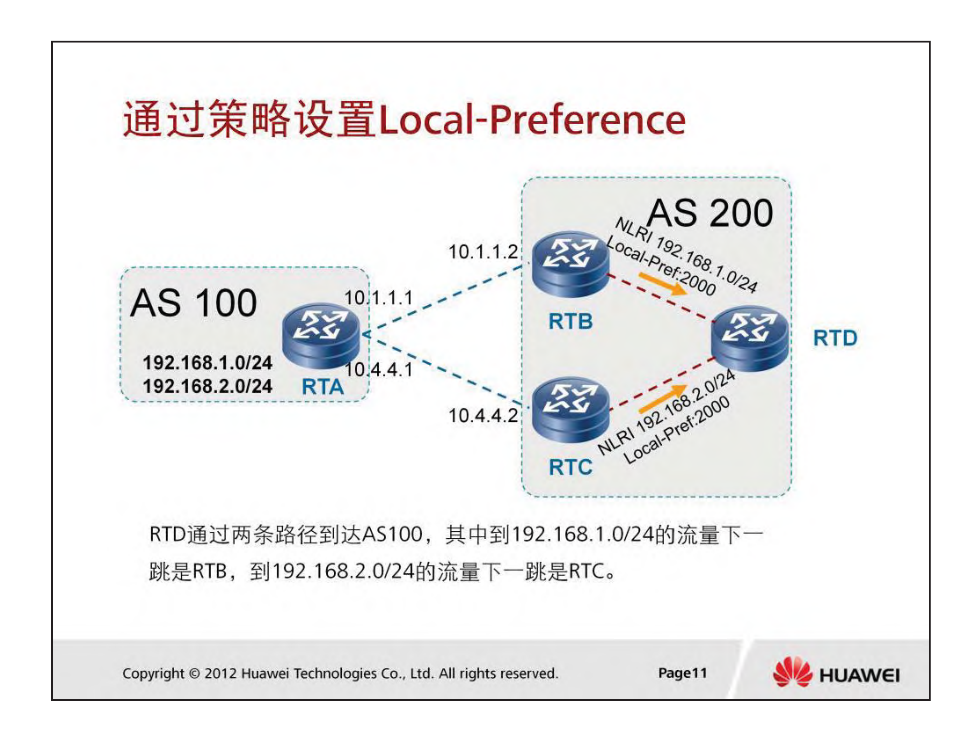
LOCAL_PREF属性通常用于影响AS内部出流量的路径选择。当AS内部的BGP Speaker存在多条到达同一外部目的地的路径时，LOCAL_PREF就变得相当有用。

default local-preference命令用来配置BGP的缺省本地优先级，该值越大则优先级越高。华为设备LOCAL_PREF数值的范围是0-4294967295，默认值为100。



如图所示，RTD将会收到两条到达同一目的地192.168.1.1/32的更新信息，分别来自RTB和RTC。默认的情况下，RTD会进行BGP路径选择，选择最佳的一条路由进行数据的转发。

RTB和RTC都通过命令“default local-preference”分别把本地默认LOCAL_PREF值修改为2000以及1000，即RTD收到一条带有LOCAL_PREF为2000的路由到达192.168.1.1/32，其下一跳为RTB，收到另外一条带有LOCAL_PREF为1000的路由到达192.168.1.1/32，其下一跳为RTC。在这情况下(其它参数都采用默认值)，下一跳为RTB的链路将会被选择为最佳路由，负责RTD到达192.168.1.1/32的数据转发。



LOCAL_PREF是影响BGP内部选路的一个很重要的参数，BGP可以结合其它一些策略工具，在一些更复杂的网络里实现负载分担。

如图所示，RTD可以通过两条路径到达AS100内的192.168.1.0/24以及192.168.2.0/24，通过策略设置Local-Preference实现到192.168.1.0/24的流量下一跳为RTB，而到达192.168.2.0/24的流量下一跳为RTC。

路由器RTB的策略配置

```
#
acl number 2000
 rule 5 permit source 192.168.1.0 0.0.0.255
#
bgp 200
 peer 10.1.1.1 as-number 100
 peer 3.3.3.3 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  peer 10.1.1.1 enable
  peer 10.1.1.1 route-policy test1 import
#
route-policy test1 permit node 10
 if-match acl 2000
 apply local-preference 2000
route-policy test1 permit node 20
 apply local-preference 1000
#
```

 Page12 

配置说明：

```
#
acl number 2000
 rule 5 permit source 192.168.1.0 0.0.0.255
#
```

指定路由匹配的范围

```
bgp 200
 peer 10.1.1.1 as-number 100
 peer 3.3.3.3 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  peer 10.1.1.1 enable
  peer 10.1.1.1 route-policy test1 import
#
```

对从对等体10.1.1.1接收的路由信息引用路由策略：test1。

```
#
route-policy test1 permit node 10
 if-match acl 2000
 apply local-preference 2000
route-policy test1 permit node 20
 apply local-preference 1000
#
```

路由策略，node 10的将匹配ACL 2000的网络分配LOCAL_PREF 2000，node 20的将其它所有不匹配ACL 2000的网络分配LOCAL_PREF 1000。

路由器RTC的策略配置

```
#
acl number 2000
 rule 5 permit source 192.168.2.0 0.0.0.255
#
bgp 200
 peer 10.1.1.1 as-number 100
 peer 3.3.3.3 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  peer 10.1.1.1 enable
  peer 10.1.1.1 route-policy test1 import
#
route-policy test1 permit node 10
 if-match acl 2000
 apply local-preference 2000
route-policy test1 permit node 20
 apply local-preference 1000
#
```




BGP MED

default med命令用来配置BGP的缺省MED值。

[Router-bgp] **default med** ***med***

缺省情况下，MED的值为0。

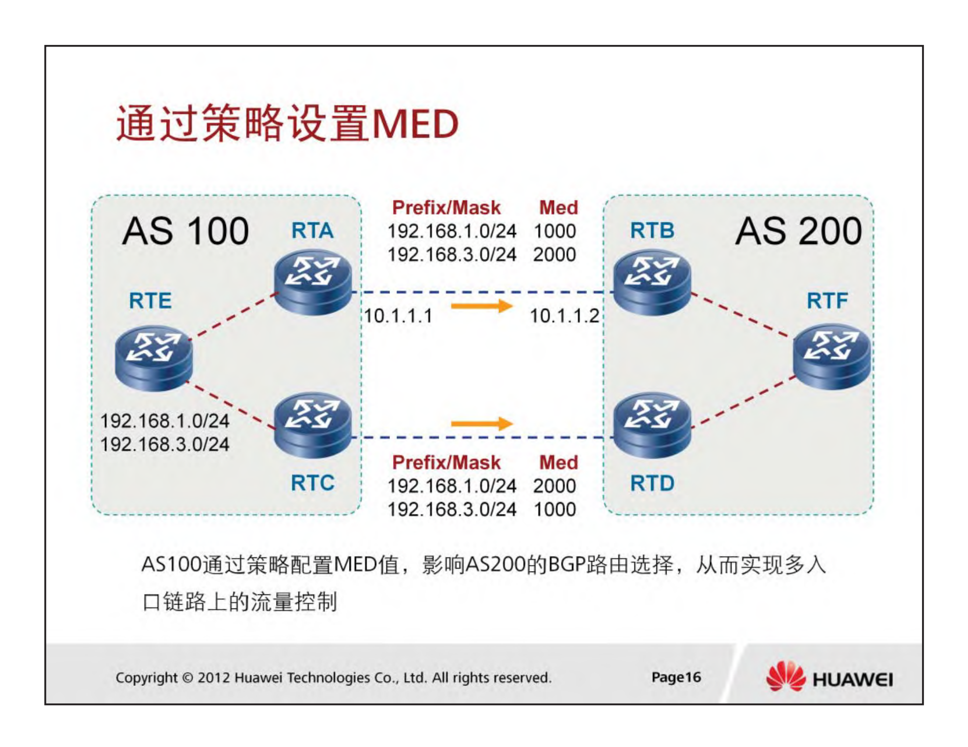
配置不同MED值会影响BGP的路由选择。

MED值越小，越优先，通常我们把MED值当作Cost来使用。

MED属性仅在相邻两个AS之间传递，收到此属性的AS不会再通告给任何其他第三方AS。



当存在到某个网络有多个出口的时候，以MED作为度量值指明最佳入口路径。MED属性相当于我们平常熟悉的COST值，取值范围0-4284967295，数值越小越优先，默认值为0。



AS100通过策略配置MED值，影响AS200的BGP路由选择，从而实现多入口链路上的流量控制。

AS100里有两个网络，192.168.1.0/24和192.168.3.0/24，分别通过RTA和RTC更新到AS200。在RTA上配置192.168.1.0/24的MED为1000，192.168.3.0/24的MED值为2000；另外在RTC上配置192.168.1.0/24的MED值为2000，而192.168.3.0/24的MED为1000。这样，从AS200内RTF到达192.168.3.0/24的流量就从RTC进来，到达192.168.1.0/24的流量就从RTA进来。在AS100内实现了基于入流量的负载分担。

路由器RTA的策略配置

```
#
bgp 100
 peer 10.1.1.2 as-number 200
 peer 3.3.3.3 as-number 100
 peer 5.5.5.5 as-number 100
 #
 ipv4-family unicast
  undo synchronization
  peer 10.1.1.2 enable
  peer 10.1.1.2 route-policy test1 export
  peer 3.3.3.3 enable
  peer 5.5.5.5 enable
#
route-policy test1 permit node 10
 if-match ip-prefix 1
 apply cost 2000
route-policy test1 permit node 20
 apply cost 1000
#
 ip ip-prefix 1 index 10 permit 192.168.3.0 24 greater-equal 24 less-equal 24
#
```

　Page17



配置说明：

```
#
bgp 100
 peer 10.1.1.2 as-number 200
 peer 3.3.3.3 as-number 100
 peer 5.5.5.5 as-number 100
 #
 ipv4-family unicast
  undo synchronization
  peer 10.1.1.2 enable
  peer 10.1.1.2 route-policy test1 export                \\引用针对出方向的路由策略test1，
  peer 3.3.3.3 enable
  peer 5.5.5.5 enable
#
```

route-policy test1 permit node 10

\\路由策略test1，node 10

if-match ip-prefix 1

\\如果匹配ip-prefix 1，则应用cost(即MED)为2000apply cost 2000

route-policy test1 permit node 20

\\路由策略test1，node 20

apply cost 1000

\\其它所有不匹配ip-prefix 1的路由都应用cost值为1000

#

ip ip-prefix 1 index 10 permit 192.168.3.0 24 greater-equal 24 less-equal 24 \\通过ip-prefix列表定义地址范围为192.168.3.0/24

#

路由器RTB的策略配置

```
#
bgp 100
 peer 10.4.4.1 as-number 200
 peer 1.1.1.1 as-number 100
 peer 5.5.5.5 as-number 100
 #
 ipv4-family unicast
  undo synchronization
  peer 10.4.4.1 enable
  peer 10.4.4.1 route-policy test1 export
  peer 1.1.1.1 enable
  peer 5.5.5.5 enable
#
route-policy test1 permit node 10
 if-match ip-prefix 1
 apply cost 2000
route-policy test1 permit node 20
 apply cost 1000
#
 ip ip-prefix 1 index 10 permit 192.168.1.0 24 greater-equal 24 less-equal 24
#
```


Page19


AS-PATH Filter

在同一个列表编号下，可以定义多条过滤规则（permit或deny）。在匹配过程中，这些规则之间是“或”的关系，即只要路由信息通过其中一项规则，就认为通过由该列表编号标识的这组AS路径过滤列表。

AS-PATH Filter 通过正则表达式过滤AS_PATH属性信息。

 Page20 

在同一个列表编号下，可以定义多条过滤规则（permit或deny）。在匹配过程中，这些规则之间是“或”的关系，即只要路由信息通过其中一项规则，就认为通过由该列表编号标识的这组AS路径过滤列表。

AS-PATH Filter 通过正则表达式过滤AS_PATH属性信息。

正则表达式

正则表达式只是BGP过滤的一种方法。

正则表达式是按照一定的规则来匹配字符串的公式。基于这些字符串对BGP路由的AS_PATH属性做出判断（接收或者拒绝）。实际上可以认为它是一个AS_PATH的ACL。

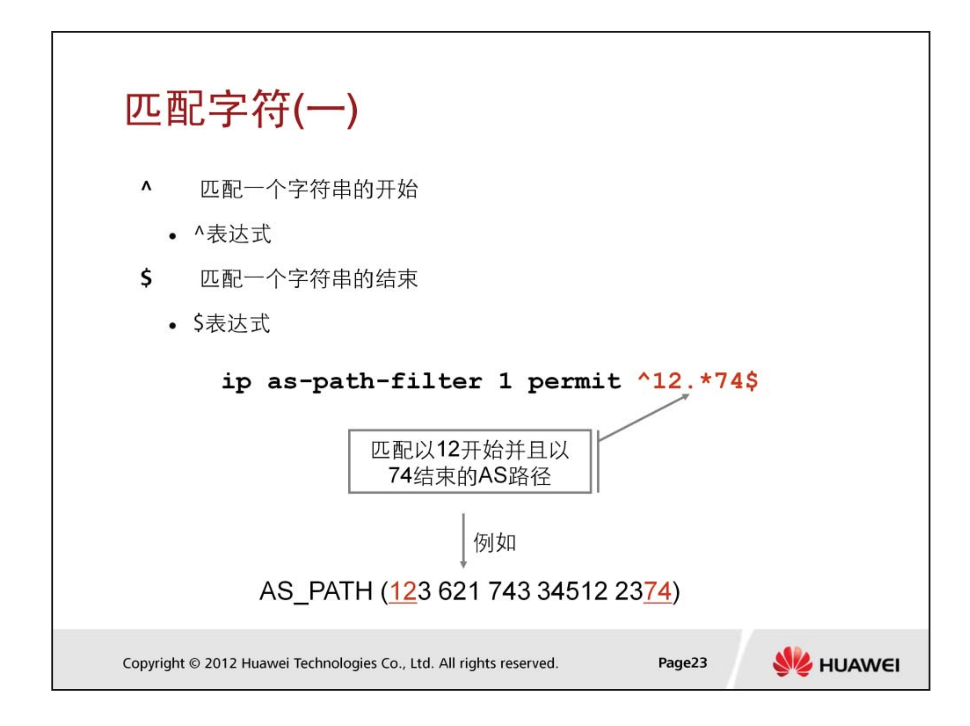
正则表达式可以定义多个permit或deny的语句，语句与语句之间是“或”的关系。



正则表达式

符号	说　明
^	匹配一个字符串的开始。如“^200”表示只匹配AS_PATH的第一个值为200。
$	匹配一个字符串的结束。如“200$”表示只匹配AS_PATH的最后一个值为200。
.	匹配任何单个字符，包括空格。
+	匹配前面的一个字符或者一个序列，可以一次或者多次出现。
_	匹配一个符号。如逗号，括号，空格符号等。
*	匹配前面的一个字符或者一个序列，可以零次或多次出现。
( )	匹配变化的AS或者一个独立的匹配，通常和“|”一起使用。
|	逻辑或。
[ ]	匹配一个范围内的AS，通常和“-”一起使用。
-	连接符。

 Page22 



"^"表示匹配字符串的开始，那么正则表达式：^12可以匹配AS_PATH列表里的第一个AS号123。换另一种说法的话，就表示该路由信息所经过的最后一个AS，AS号必须是以12开始。

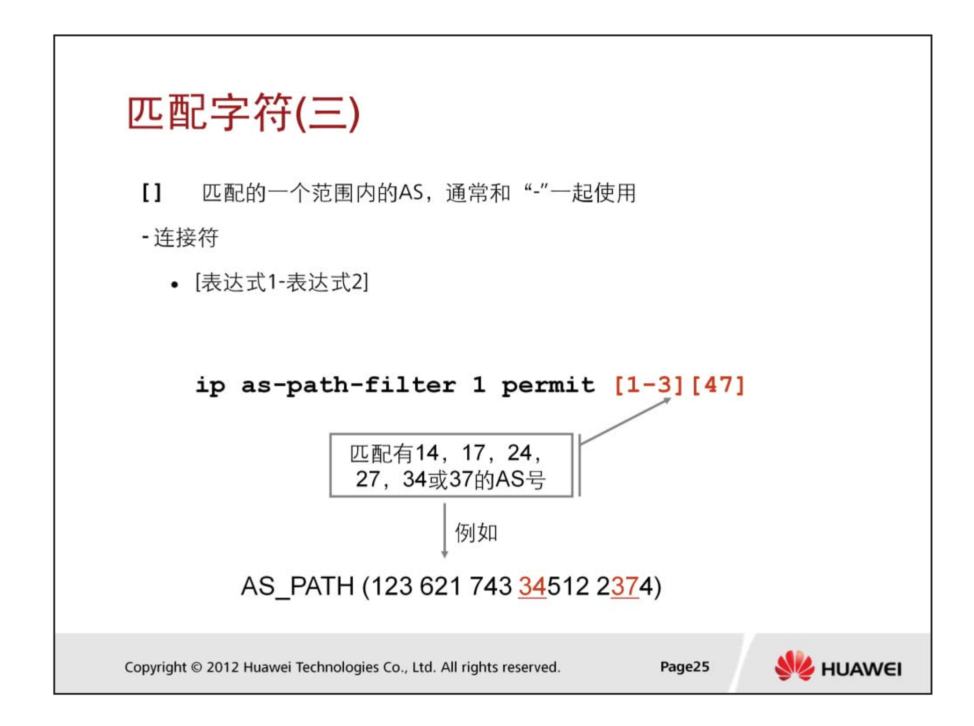
"$"表示匹配字符串的结束，正则表达式：74$可以匹配AS_PATH列表里的最后一个AS号2374，即路由信息所经过的第一个AS的AS号必须是以74结尾。

正则表达式必须有始有终。"^"在一串匹配符前必须排在开始端。

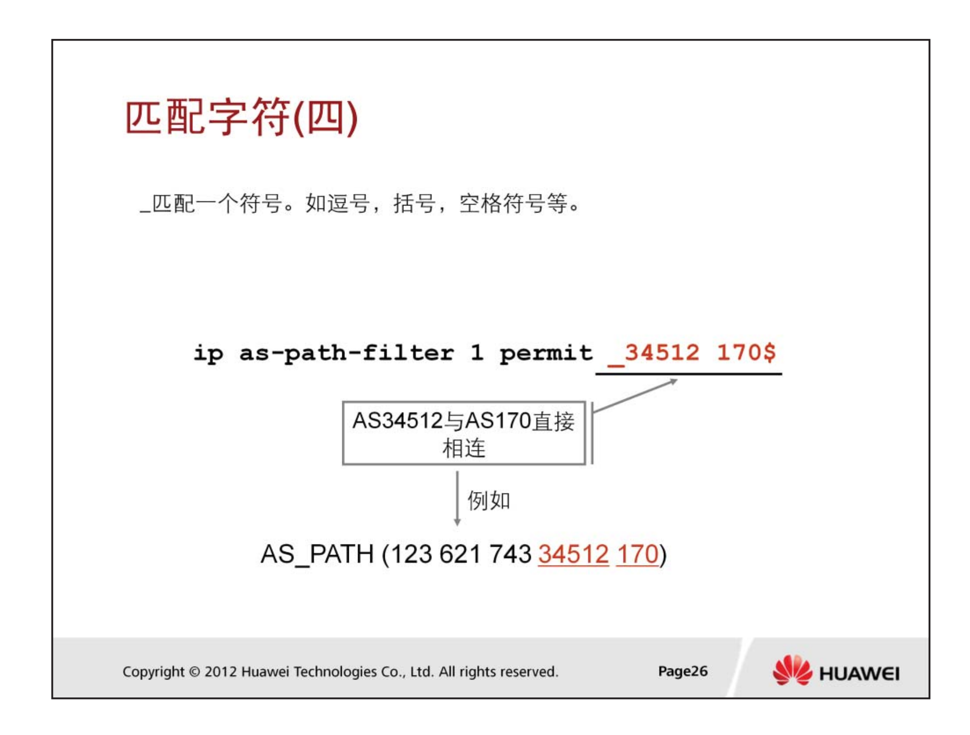


正则表达式：23|43 匹配在AS_PATH列表里存在23或43的AS。

如图所示，该AS_PATH属性的一串AS号与23|43可以匹配3次，“123 621 743 34512 2374”。正则表达式是可以匹配一个AS号里的某些字符，如匹配“743”里的“43”，匹配“123”和“2374”里的23。



正则表达式[1-3]表示匹配“1”，“2”或“3”的字符，而正则表达式[47]则表示匹配4或者7的字符，当两个正则表达式共同使用时，则表示匹配有14,17,24,27,34或37的AS号。



“.”与“_”的区别在于“.”可以匹配字符以及符号，其中还包括空格，而“_”只能匹配一个符号，如逗号，括号空格等。

“_34512 170$”表示AS34512和AS170直接相连。这里“_”表示一个符号，可以匹配AS_PATH (123 621 743 34512 170)中“743 34512”之间的空格。

匹配字符(五)

. 匹配任何单个字符，包括空格。

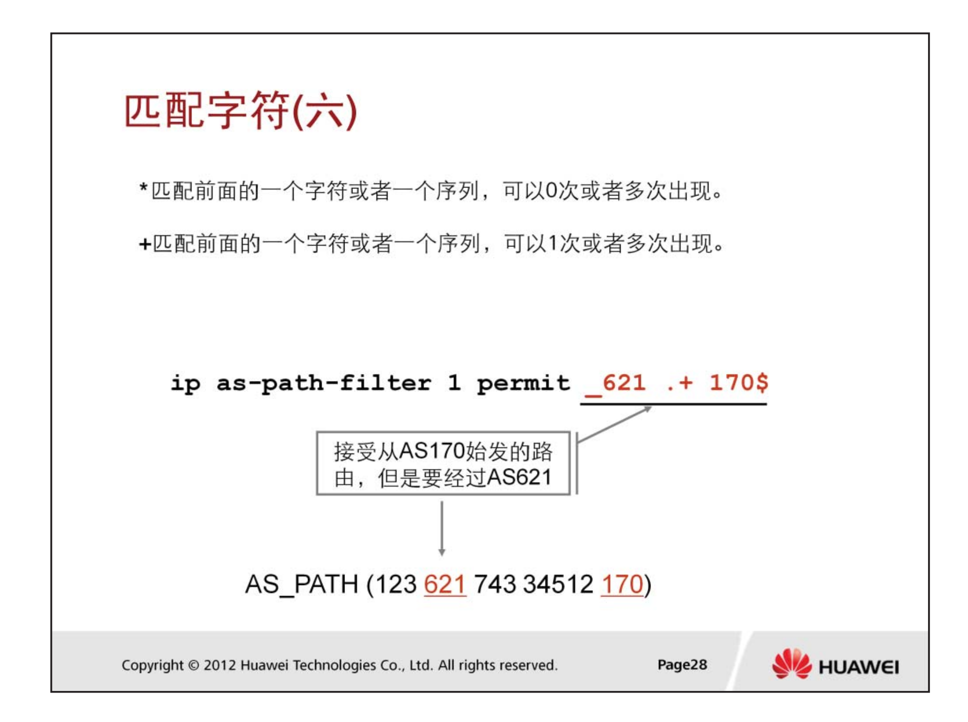
```
ip as-path-filter 1 permit [1-3].[47]
```

AS_PATH (123 621 743 34512 2374)

AS_PATH (123 621 743 34512 2374)

Page27 HUAWEI

[1-3].[47]由于中间的“.”可以匹配任何一个字符或符号，所以在AS_PATH (123 621 743 34512 2374)列表里总共可以匹配3次，分别是“1 7”，“237”，“374”。这里需要注意，“.”可以匹配空格。



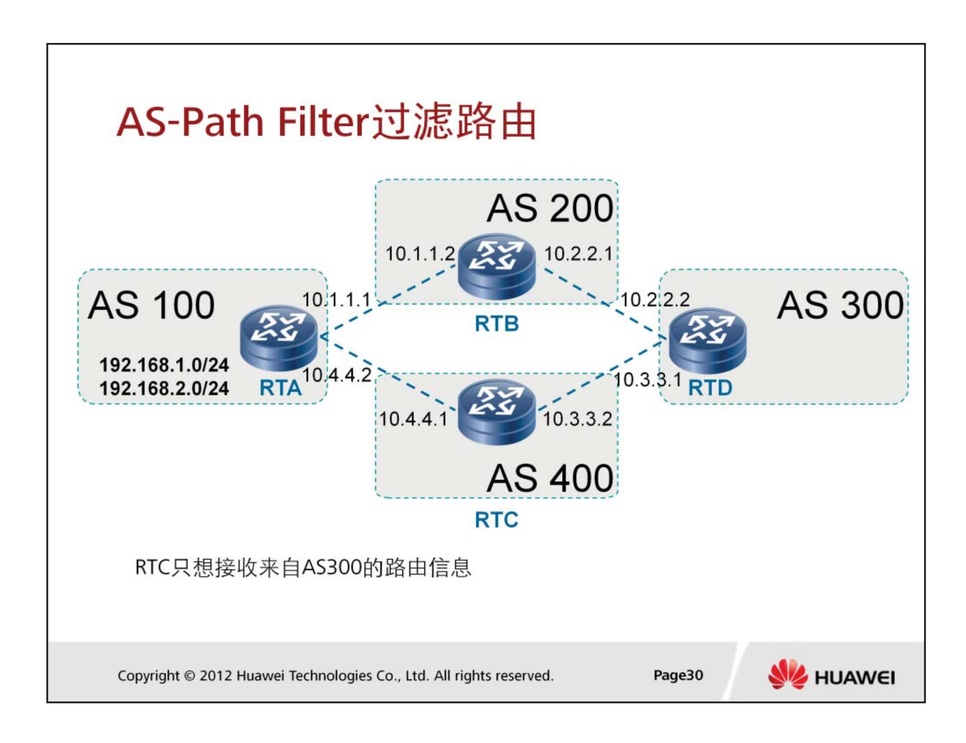
如上图所示，相信难点在于中间的“.+”。回顾一下，正则表达式里“.”表示任何一个字符或符号，而“+”表示匹配前面的一个字符或一个序列，可以1次或多次出现。那“.+”则可以表示为多个任意字符或符号，所以在该例子中匹配了AS_PATH (123 621 743 34512 170)列表中的“ 743 34512 ”。

一些常用的正则表达式

正则表达式	涵 义
^$	匹配本地AS始发的路由
.*	匹配所有路由
_10_	匹配所有必须通过AS10的路由
^10$	匹配AS-PATH中只有AS10的路由
^10_.	匹配从相邻AS10接收的路由
^[0-9]+$	AS_PATH只有一个AS号

Page29





按照图中的要求，可以知道只需在RTC上对两个EBGP对等体(RTA和RTD)实施入方向的过滤策略就可实现。

路由器RTC的配置

```
#
bgp 400
 peer 10.4.4.2 as-number 100
 peer 10.3.3.1 as-number 300
 #
 ipv4-family unicast
  undo synchronization
  peer 10.4.4.2 enable
  peer 10.4.4.2 as-path-filter 1 import
  peer 10.3.3.1 enable
  peer 10.3.3.1 as-path-filter 1 import
#
 ip as-path-filter 1 permit ^300_
#
```

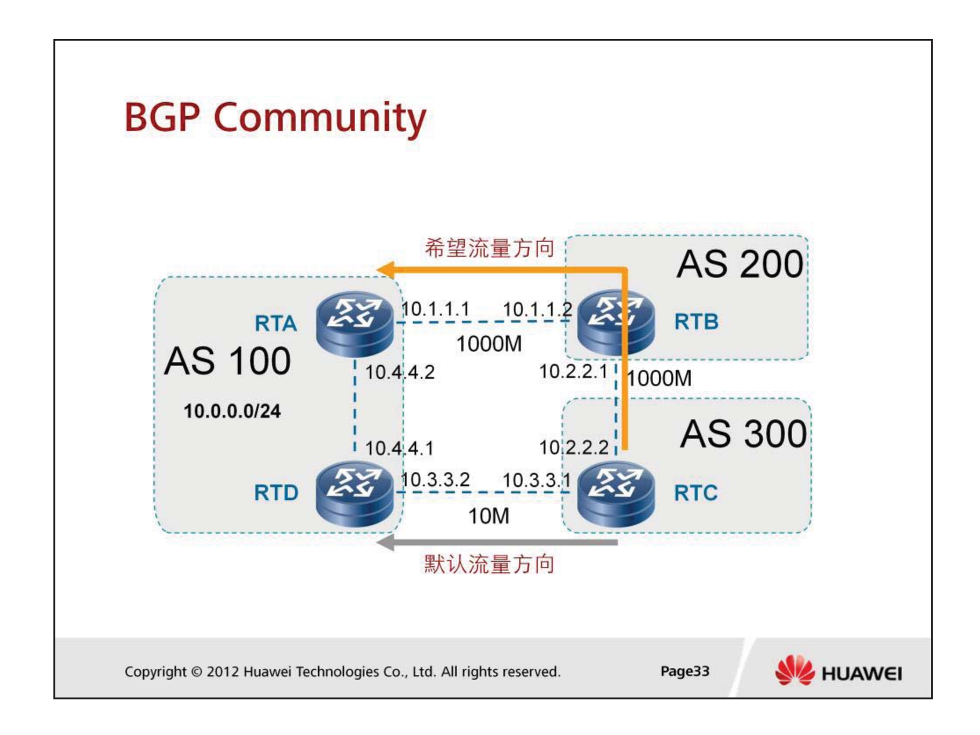

Page31



配置说明:

```
#
bgp 400
 peer 10.4.4.2 as-number 100
 peer 10.3.3.1 as-number 300
 #
 ipv4-family unicast
  undo synchronization
  peer 10.4.4.2 enable
  peer 10.4.4.2 as-path-filter 1 import                    \\对RTA应用入方向的as-path-filter
  peer 10.3.3.1 enable
  peer 10.3.3.1 as-path-filter 1 import                    \\对RTD应用入方向的as-path-filter
#
```

```
ip as-path-filter 1 permit ^300_
                              \\只接收来自AS300的路由信息
#
```



在图中的例子里，AS100里的两ASBR宣告网络10.0.0.0/24到网络里所有节点。RTB和RTC都会有两条路径到达AS100，在默认的情况下，路由器会选择最优的路径到达AS100，如：RTC去往10.0.0.0/24的数据流量就会从RTC与RTD之间的链路到达AS100。但从上图可以看出，RTC与RTD之间的链路带宽只有10M，而RTA，RTB，RTC之间的链路带宽都是1000M，现希望选择从RTC-RTB-RTA的路径到达网络10.0.0.0/24。

路由器RTA的配置

```
bgp 100
 peer 10.4.4.1 as-number 100
 peer 10.1.1.2 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  peer 10.4.4.1 enable
  peer 10.1.1.2 enable
  peer 10.1.1.2 route-policy set_community export
  peer 10.1.1.2 advertise-community
#
route-policy set_community permit node 10
 apply community 100:1
```

 Page34



配置说明：

bgp 100

 peer 10.4.4.1 as-number 100

 peer 10.1.1.2 as-number 200

 #

 ipv4-family unicast

  undo synchronization

  peer 10.4.4.1 enable

  peer 10.1.1.2 enable

  peer 10.1.1.2 route-policy set_community export

\\在出方向上应用路由策略set_community

  peer 10.1.1.2 advertise-community

\\用来配置将团体属性发布给对等体(组)，缺省情况下，不将团体属性发布给任何对等体（组）

#

route-policy set_community permit node 10

apply community 100:1

\\ 设置团体属性为 100:1

路由器RTD的配置

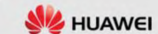
```
bgp 100
 peer 10.4.4.2 as-number 100
 peer 10.3.3.1 as-number 300
 #
 ipv4-family unicast
  undo synchronization
  peer 10.4.4.2 enable
  peer 10.3.3.1 enable
  peer 10.3.3.1 route-policy set_community export
  peer 10.3.3.1 advertise-community
#
route-policy set_community permit node 10
 apply community 100:2
```


Page36



路由器RTC的配置

```
bgp 300
 peer 10.2.2.1 as-number 200
 peer 10.3.3.2 as-number 300
 #
 ipv4-family unicast
  undo synchronization
  peer 10.2.2.1 enable
  peer 10.2.2.1 route-policy set_local_pref import
  peer 10.2.2.1 advertise-community
  peer 10.3.3.2 enable
  peer 10.3.3.2 route-policy set_local_pref import
  peer 10.3.3.2 advertise-community
#
route-policy set_local_pref permit node 10
 if-match community-filer 1
 apply local-preference 200
Route-policy set_local_pref permit node 20
 if-match community-filter 2
 apply local-preference 50
#
 ip community-filter 1 permit  100:1
 ip community-filter 2 permit  100:2
```

 Page37 

配置说明:

#

 ip community-filter 1 permit 100:1 \\配置团体属性过滤列表

 ip community-filter 2 permit 100:2

#

ip community-filter basic-comm-filter-num { deny | permit } [community-number | aa:nn] * &<1-16> [internet | no-export-subconfed | no-advertise | no-export]

ip community-filter adv-comm-filter-num { deny | permit } regular-expression

在基本团体属性列表只能指定团体号或团体属性（ basic-comm-filter-num 的取值范围为1～99），

在高级团体属性列表中则可以使用正则表达式作为匹配条件（ adv-comm-filter-num 的取值范围为100~199）。

例：

配置序号为1的基本团体属性列表。

[Quidway] ip community-filter 1 permit internet

配置序号为100的高级团体属性列表。

[Quidway] ip community-filter 100 permit ^10

查看团体属性

```
[RTC]display bgp routing-table community

 Total Number of Routes: 2

 BGP Local router ID is 10.2.2.2
 Status codes: * - valid, > - best, d - damped,
               h - history,  i - internal, s - suppressed, S - Stale
               Origin : i - IGP, e - EGP, ? - incomplete
      Network            NextHop        MED        LocPrf    PrefVal Community

 *    10.0.0.0/24        10.3.3.2       0           50         0      <100:2>
 *>                      10.2.2.1                   200        0      <100:1>
```


HUAWEI

display bgp routing-table community [aa:nn &<1-13>] [no-advertise | no-export | no-export-subconfed][whole-match]

community：用来查看BGP路由表中指定团体的路由信息。

aa:nn：指定的团体号。

no-advertise：显示带有No-Advertise团体属性的BGP路由。

no-export：显示带有No-Export团体属性的BGP路由。

no-export-subconfed：显示带有No-Export-Subconfed团体属性的BGP路由。

whole-match：精确匹配。

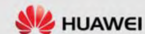
查看团体属性 (续)

```
[RTC]display bgp routing-table 10.0.0.0

 BGP local router ID : 10.2.2.2
 Local AS number : 300
 Paths:   2 available, 1 best

 BGP routing table entry information of 10.0.0.0/24:
 From: 10.2.2.1 (10.1.1.2)
 Original nexthop: 10.2.2.1
 Community:<100:1>
 AS-path 200 100, origin igp, localpref 200, pref-val 0, valid, external, best,
pre 255
 Advertised to such 1 peers:
    10.3.3.2

 BGP routing table entry information of 10.0.0.0/24:
 From: 10.3.3.2 (10.3.3.2)
 Original nexthop: 10.3.3.2
 Community:<100:2>
 AS-path 100, origin igp, MED 0, localpref 50, pref-val 0, valid, external, pre
255
 Not advertised to any peer yet
```

 Page40 

display bgp routing-table [ipv4-address] [{ mask | mask-length } [longer-prefixes]]

ipv4-address：网络地址。

mask/mask-length：点分十进制掩码/掩码长度。

longer-prefixes：允许按更长的掩码匹配。



问 题

BGP路由策略工具有哪些?

影响BGP路由选择的参数有哪些?

请说出在正则表达式里，“+”与“*”的区别？“.”与“_”的区别？

Page41



1.BGP路由策略工具有哪些?

主要有ACL，IP-PREFIX列表，Filter-List，路由策略以及专门为BGP设计的AS-PATH-FILTER和COMMUNITY-FILTER。

2.影响BGP路由选择的参数有哪些?

影响BGP选路的重要参数主要有以下这些:

- Preferred Value
- Local-Preference
- AS-Path
- Origin
- MED
- EBGP/IBGP
- IGP Cost
- CLUSTER ID
- Communities

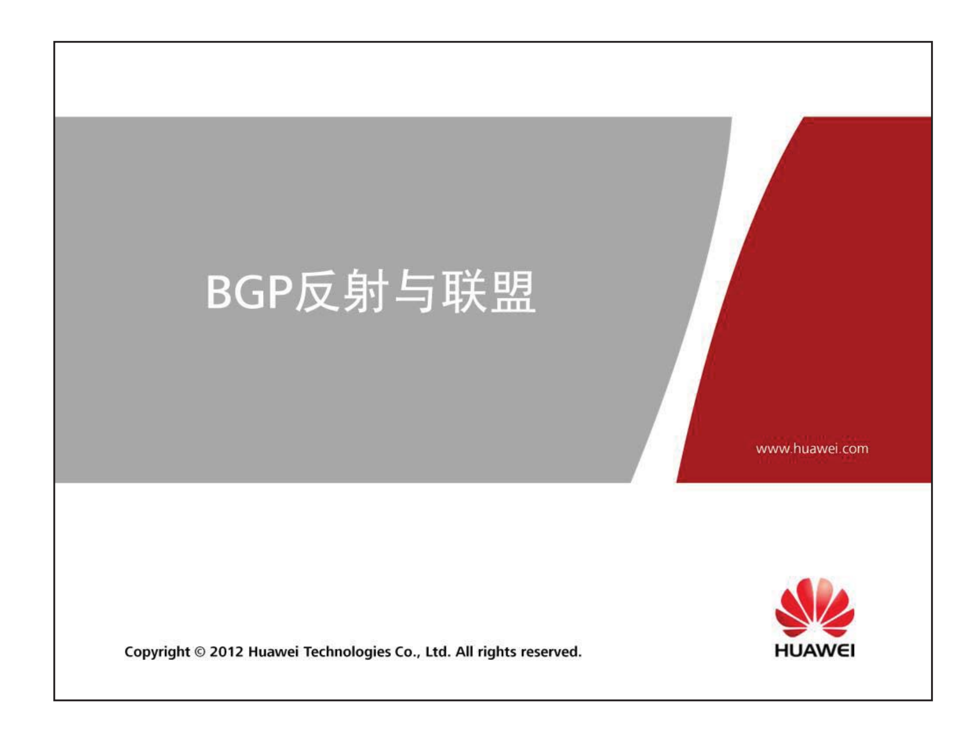
其中我们常用的有Preferred Value，Local-Pref，AS_PATH，MED和Community。

3.请说出在正则表达式里，“+”与“*”的区别？“.”与“_”的区别？

“+”号表示1次或多次匹配前面的字符；而“*”号表示0次或多次匹配前面的字符。

“.”号表示任意一个字符，其中还包括空格；而“_”号表示一个符号，其中有空格、逗号、括号等AS_PATH里的符号。


谢谢
www.huawei.com


BGP反射与联盟
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

为了实现路由信息的交互，BGP要求一个AS内的所有BGP Speaker相互形成IBGP对等体全互连，而这一要求使得IBGP的扩展成为了一个很大的问题。BGP反射器以及BGP联盟则是为了解决该问题而提出的BGP扩展技术。





培训目标

学完本课程后，您应该能：

- 知道BGP反射器的工作原理
- 知道BGP联盟的工作原理

 Page2





目 录

路由反射和联盟简介

BGP路由反射

BGP联盟

Page3







目 录

路由反射和联盟简介

BGP路由反射

BGP联盟



IBGP扩展性的问题

BGP是怎样防止环路的?

- EBGP
  - 通过AS-Path属性，丢弃从EBGP对等体接收到的在AS-Path属性里包含自身AS号的任何更新信息
- IBGP
  - BGP路由器不会将任何从IBGP对等体接收到的更新信息传给其它IBGP对等体

 Page5 

IBGP扩展性的问题

IBGP防止环路机制带来的问题

- 为保证更新信息可以到达所有IBGP对等体
  - 解决方案：IBGP Speaker与IBGP Speaker之间要保证会话的全互连
    - 从而又带来**IBGP**会话数**n(n-1)/2**的问题
- 路由反射 (RFC2796)
- 联盟 (RFC3065)

Page6



IBGP扩展问题解决方案

路由反射 (RFC 2796)

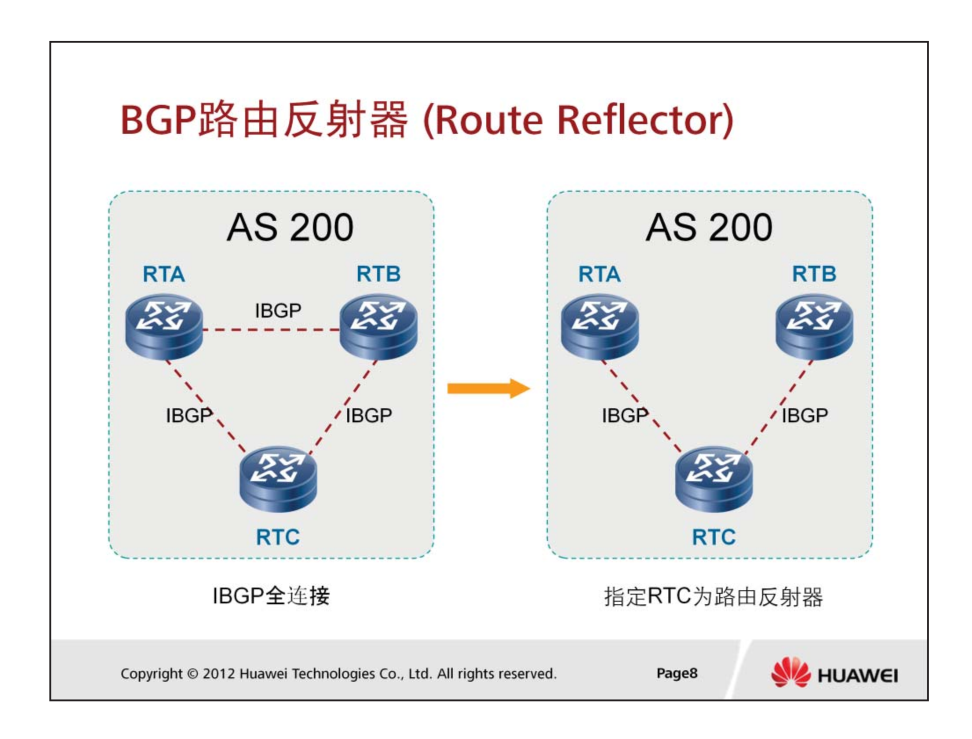
- 降低对指定路由器IBGP路由通告机制的限制，允许将从IBGP对等体接收到的更新信息传给某些IBGP对等体

联盟 (RFC3065)

- 将大的AS分成若干小的AS，而小AS之间建立EBGP对等体关系

Page7

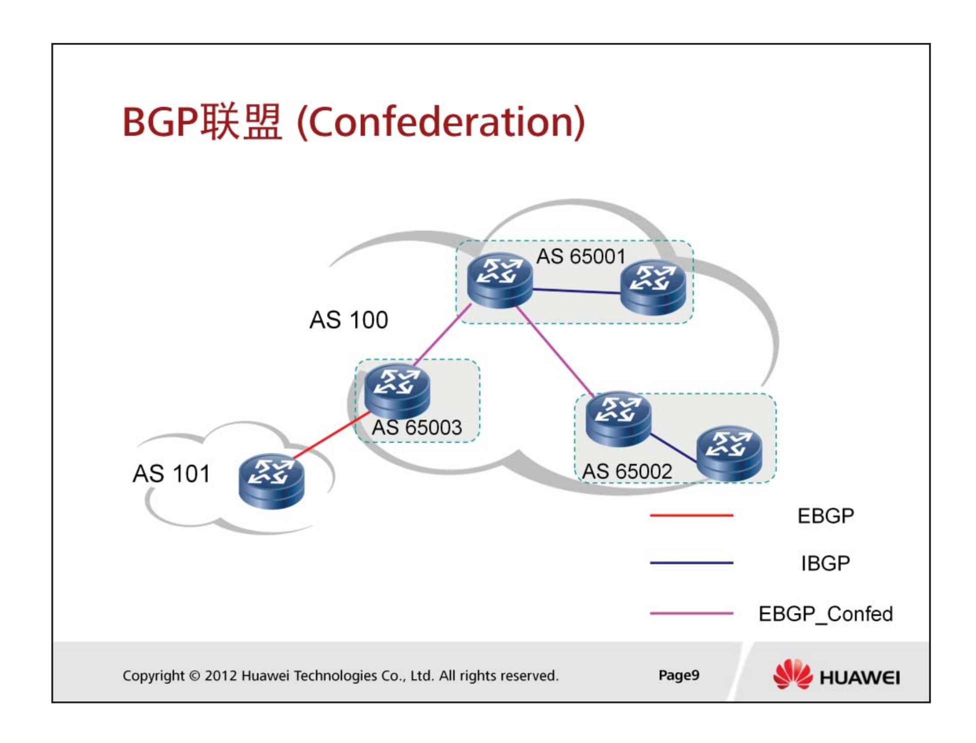




在AS200里，有三台路由器分别为RTA，RTB和RTC。在默认的情况下，如果RTA收到一条外部的路由更新，并且该路由被RTA选举为最佳路由，则RTA肯定会把该路由通告给RTB以及RTC。由于RTB和RTC互为IBGP对等体，所以不会把从IBGP学习到的路由通告给其它IBGP对等体。

如果该通告原则可以被放松，允许RTC可以把从RTA学习到的IBGP路由通告给其它IBGP对等体的话，这样将可以取消RTA与RTB之间的IBGP会话。

RTC就是BGP路由反射器。



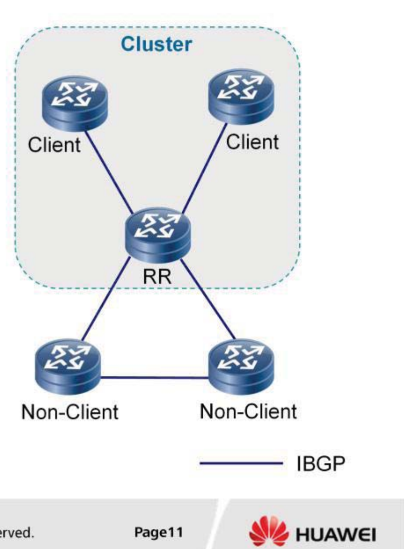
联盟通过把大的AS分成多个更小的自治系统来解决IBGP全互连的问题，这些自治系统叫做成员自治系统或子自治系统。成员自治系统之间使用EBGP会话，因此它们不需要全互连。然而，在每一个成员AS中，仍然要求IBGP全互连。


目 录
路由反射和联盟简介
BGP路由反射
BGP联盟
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page10
HUAWEI

不同角色的对等体

IBGP对等体可以有三种角色:

- 路由反射器 (Route Reflector)
- 客户机 (Client)
- 非客户机 (Non-Client)



Page11
HUAWEI

我们使用路由反射来描述一个BGP Speaker通告一条IBGP路由到另外一个IBGP对等体的操作。而这样的一个BGP Speaker通常被称为路由反射器(Route Reflector, RR)，这样的一条IBGP路由被称为反射路由。

IBGP对等体可以有三种角色:

- 路由反射器 (Route Reflector)
- 客户机 (Client)
- 非客户机 (Non-Client)
- 路由反射器和它的客户机组成一个集群（Cluster）。路由反射器在客户机之间传递（反射）路由信息，所以客户机之间不需要建立BGP连接。
- 既不是反射器也不是客户机的BGP路由器被称为非客户机（Non-Client）。非客户机与路由反射器之间，以及所有的非客户机之间仍然必须建立全连接关系。

反射器(RR)的内部对等体被分为两组:

1) 客户对等体 (Client peers)

2) 非客户对等体 (Non-Client peers)

一个RR负责在客户对等体组之间反射这些组的路由信息。一个RR和它的客户对等体形成一个单独的簇。什么是簇我们会在后面的胶片中描述。非客户对等体组之间必须IBGP全连接，而客户对等体组之间不需要建立IBGP全连接，只需要维护与RR之间的IBGP会话就足够了。如上图所示。

对等体之间的关系

Client只需维护与RR之间的IBGP会话

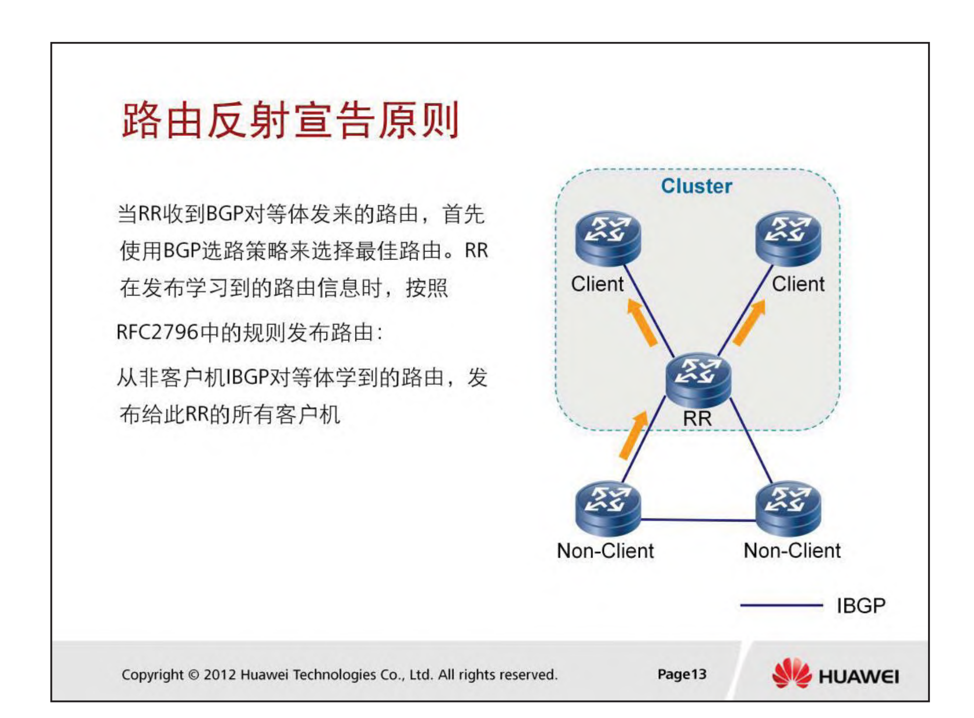
RR与RR之间需要建立IBGP的全互连

Non-Client与Non-Client之间需要建立IBGP全互连

RR与Non-Client之间需要建立IBGP全互连

Page12





当RR收到BGP对等体发来的路由，首先使用BGP选路策略来选择最佳路由。在发布学习到的路由信息时，RR按照RFC2796中的规则发布路由。

从非客户机IBGP对等体学到的路由，发布给此RR的所有客户机。

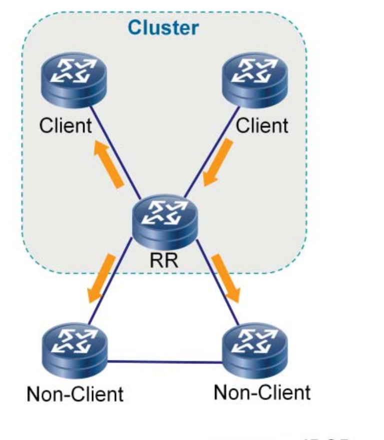
从客户机学到的路由，发布给此RR的所有非客户机和客户机（发起此路由的客户机除外）。

从EBGP对等体学到的路由，发布给所有的非客户机和客户机。

路由反射宣告原则(续)

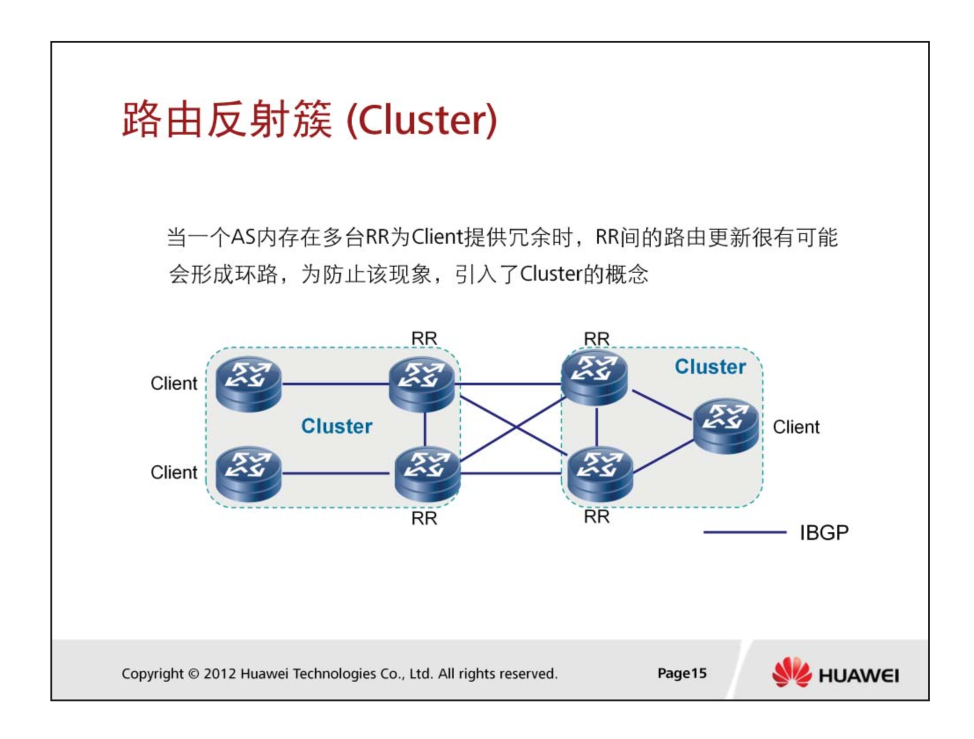
从客户机学到的路由，发布给此RR的所有非客户机和客户机（发起此路由的客户机除外）

从EBGP对等体学到的路由，发布给所有的非客户机和客户机

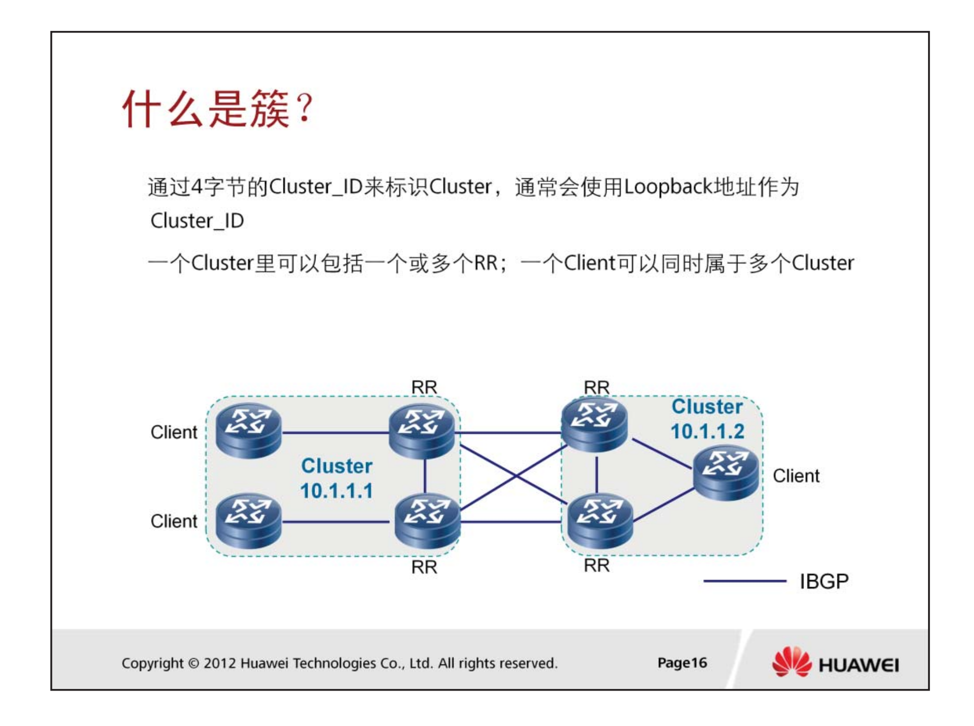




HUAWEI



当一个AS内存在多台RR为Client提供冗余时，RR间的路由更新很有可能会形成环路，为防止该现象，引入了Cluster的概念。



通常，一个客户的簇只拥有一个RR，并由RR的BGP Router-id去标识该簇。有时，为了防止单点失效，在单一簇里引入多个RR，如图中的备份RR组网。

路由反射环路防止机制－Originator_ID

Originator_ID属性用于防止在反射器和客户机/非客户机之间产生环路

Originator_ID属性长4字节，可选非过渡属性，属性类型为9，是由路由反射器（RR）产生的，携带了本地AS内部路由发起者的Router ID

当一条路由第一次被RR反射的时候，RR将Originator_ID属性加入到这条路由，标识这条路由的始发路由器。如果一条路由中已经存在了Originator_ID属性，则RR将不会创建新的Originator_ID

当其它BGP Speaker接收到这条路由的时候，将比较收到的Originator_ID和本地的Router ID，如果两个ID相同，BGP Speaker会忽略掉这条路由，不做处理

　Page17



当一个路由被反射的时候，可能会由于错误的配置形成路由环路。因此路由反射定义了以下的方法去检测和防止路由信息环路的产生：

Originator_ID属性用于防止路由在反射器和客户机/非客户机之间产生环路。

Originator_ID属性长4字节，可选非过渡属性，属性类型为9，是由路由反射器（RR）产生的，携带了本地AS内部路由发起者的Router ID。

当一条路由第一次被RR反射的时候，RR将Originator_ID属性加入这条路由，标识这条路由的始发路由器。如果一条路由中已经存在了Originator_ID属性，则RR将不会创建新的Originator_ID。

当其它BGP Speaker接收到这条路由的时候，将比较收到的Originator_ID和本地的Router ID，如果两个ID相同，BGP Speaker会忽略掉这条路由，不做处理。

路由反射环路防止机制—Cluster_List

- Cluster_List属性用于防止AS内部的环路
- Cluster_List是可选非过渡属性，属性类型编码为10
- Cluster_List由一系列的Cluster_ID组成，描述了一条路由所经过的反射器路径，这和描述路由经过的As路径的AS_Path属性有相似之处，Cluster_List由路由反射器产生
- 当RR在它的客户机之间或客户机与非客户机之间反射路由时，RR会把本地Cluster_ID添加到Cluster_List的前面。如果Cluster_List为空，RR就创建一个
- 当RR接收到一条更新路由时，RR会检查Cluster_List。如果Cluster_List中已经有本地Cluster_ID，丢弃该路由；如果没有本地Cluster_ID，将其加入Cluster_List，然后反射该更新路由

 Page18



对于AS之间，BGP用于防止环路的主要措施是通过AS_Path属性记录途经的AS路径，带有本地AS号的路由将被路由器丢弃；对于AS之内，BGP防止路由环路的方法是禁止IBGP对等体发布从AS内部学来的路由。

路由反射器的实现是基于放宽对“BGP在AS内学到的路由不会在AS中转发”的要求，即允许IBGP对等体之间发布从AS内部学来的路由。在这种情况下，Cluster_List属性被引入，用于防止AS内部的环路。

Cluster_List是可选非过渡属性，属性类型编码为10。

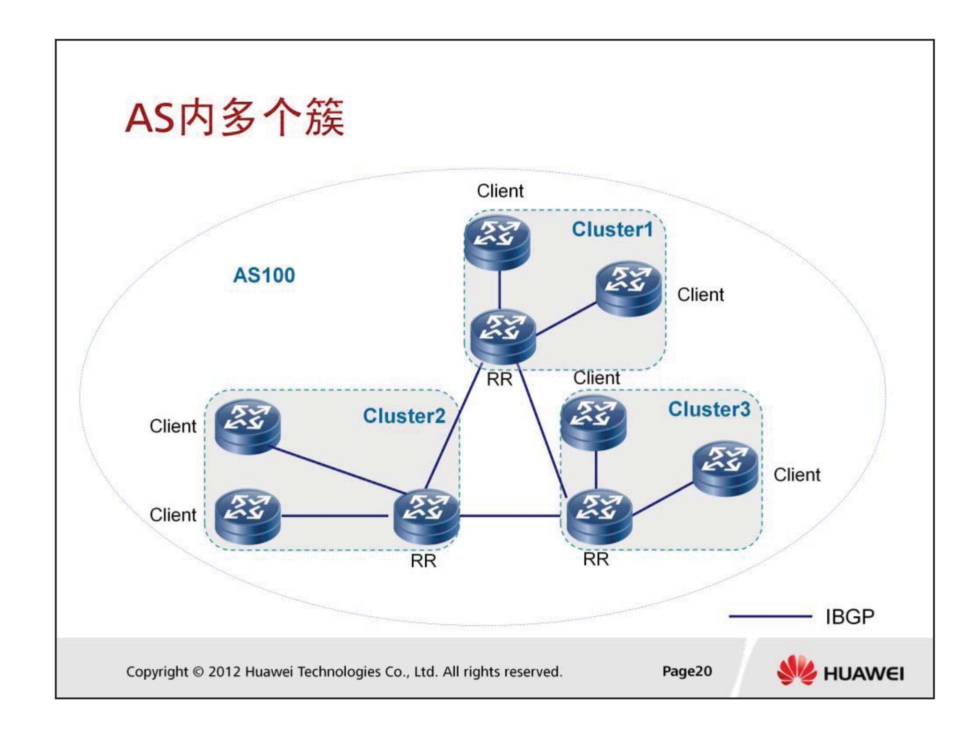
Cluster_List由一系列的Cluster_ID组成，描述了一条路由所经过的反射器路径，这和描述路由经过的As路径的AS_Path属性有相似之处。Cluster_List由路由反射器产生。

当RR在它的客户机之间或客户机与非客户机之间反射路由时，RR会把本地Cluster_ID添加到Cluster_List的前面。如果Cluster_List为空，RR就创建一个。

当RR接收到一条更新路由时，RR会检查Cluster_List。如果Cluster_List中已经有本地Cluster_ID，丢弃该路由；如果没有本地Cluster_ID，将其加入Cluster_List，然后反射该更新路由。

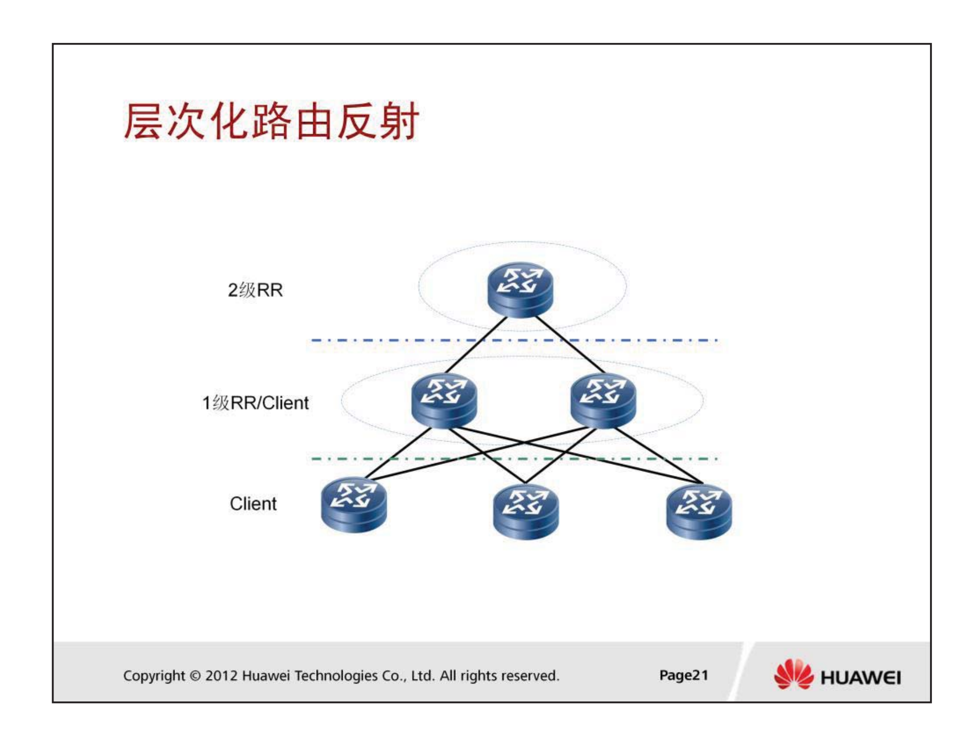
Cluster_List只被RR用来检测路由环路，不是RR的客户机和非客户机不会检测该属性。

Cluster_List只在AS内部传播，从EBGP对等体收到的含有Cluster_List的路由将被丢弃。



一个AS中可能存在多个簇（Cluster）。各个RR之间是IBGP对等体的关系，一个RR可以把另一个RR配置成自己的客户机或非客户机。因此可以灵活的配置AS内部簇与簇之间的关系。

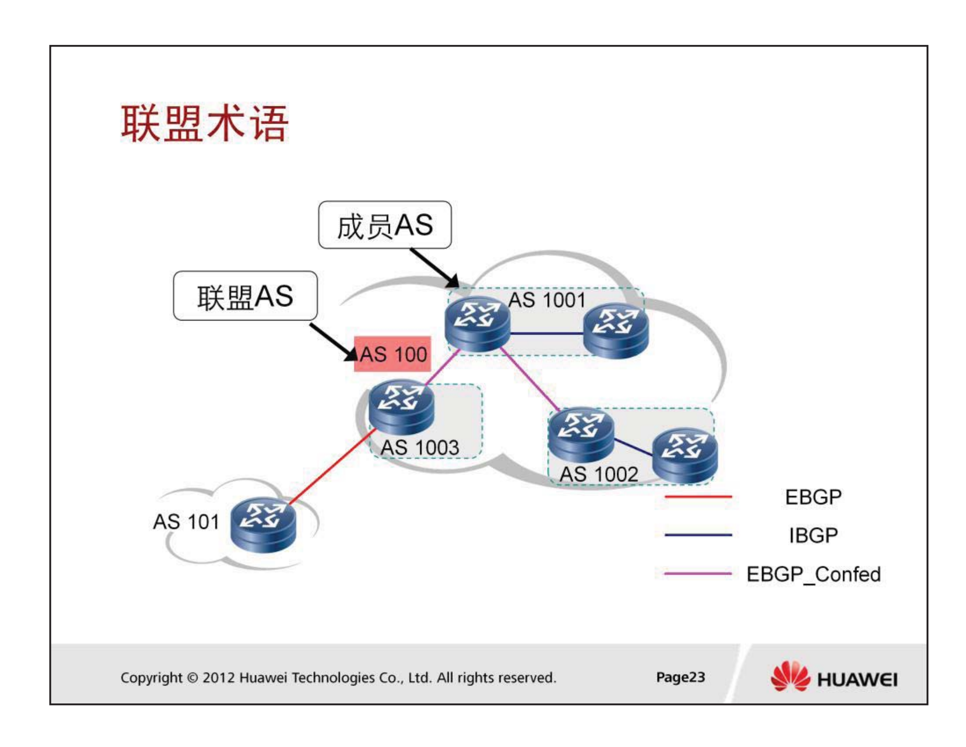
如图，一个AS内被分成多个反射簇，每个RR将其它的RR配置成非客户机，各RR之间建立全连接。每个客户机只与所在簇的RR建立IBGP连接。这样该自治系统内的所有BGP路由器都会收到反射路由信息。



路由反射减少了域中IBGP会话的总数。然而，因为RR相互之间必须全互连，在大型网络中，存在一种可能性，即RR之间仍然需要大量的IBGP会话。为了进一步减少会话数量，引入层次化的路由反射。

层次化路由反射的层数可以按照需要逐步加深，但通常在现网中两层或三层已经足够了。


目 录
路由反射和联盟简介
BGP路由反射
BGP联盟
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page22
HUAWEI

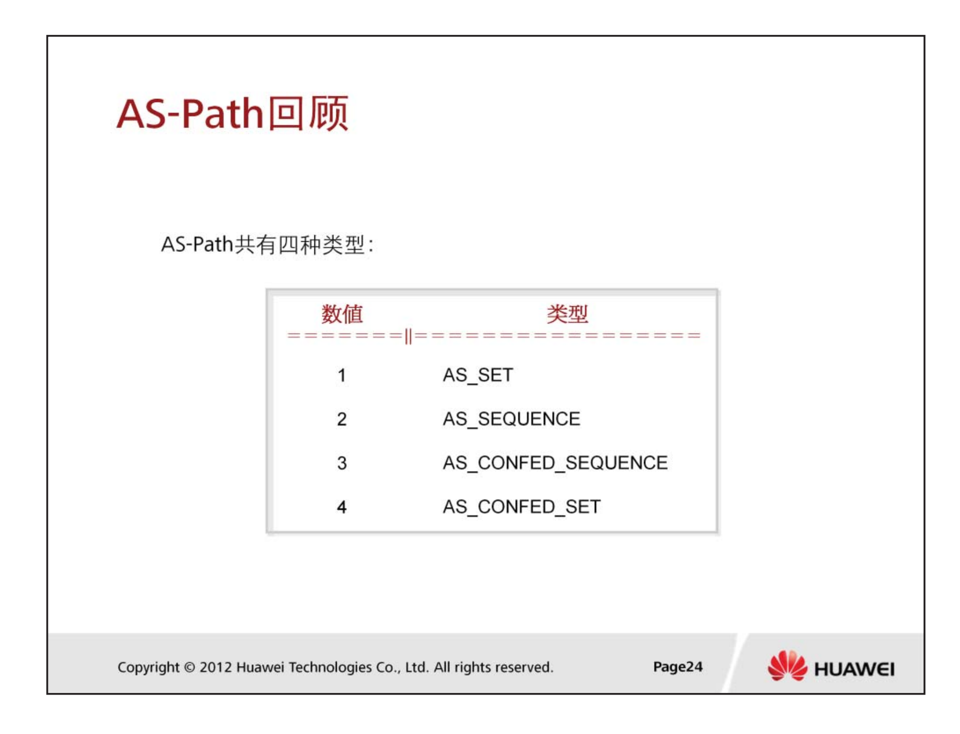


联盟通过把大的AS分成多个更小的AS来解决IBGP全互连的问题，这些自治系统叫做成员AS。因为成员AS之间使用EBGP会话，它们之间不需要全互连。然而在每一个成员AS中，IBGP全互连的要求仍然适用。

联盟中的EBGP会话和常规的EBGP会话有所不同。为了区分它们，这种类型的EBGP会话叫做联盟内的EBGP会话。与普通EBGP会话区别就发生在当通过会话传播路由的时候，联盟内的EBGP会话在一方面遵循路由通告的部分IBGP规则，在另一方面又遵循路由通告的部分EBGP规则。如：在发送更新的时候，NEXP_HOP、MED和LOCAL_PREF被保留，而AS-PATH被修改。

对于外部邻居来说(联盟外的的对等体)，成员AS拓扑是不可见的。也就是说，在发向EBGP邻居的更新消息中，已经剥去了联盟内被修改的AS_PATH。从其他的自治系统来看，联盟就像单个AS一样。

每个成员AS中，IBGP全连接是需要的。路由反射也可以被部署。部署联盟的一个明显优势就是其成员AS不需要使用相同的IGP。每个成员AS不需要向其他成员AS通告自己的内部拓扑。不过，当使用不同的IGP时，每一个成员AS内必须保证BGP下一跳的可达性。



当前，AS_PATH属性被定义为公认必遵属性，该属性由AS号所组成。AS_PATH属性字段由三元组所组成:

Path Segment Type, Path Segment Length, Path Segment Value

在BGPv4里，path segment type字段是由1字节长的数值所组成，主要是标识AS_PATH的不同类型:

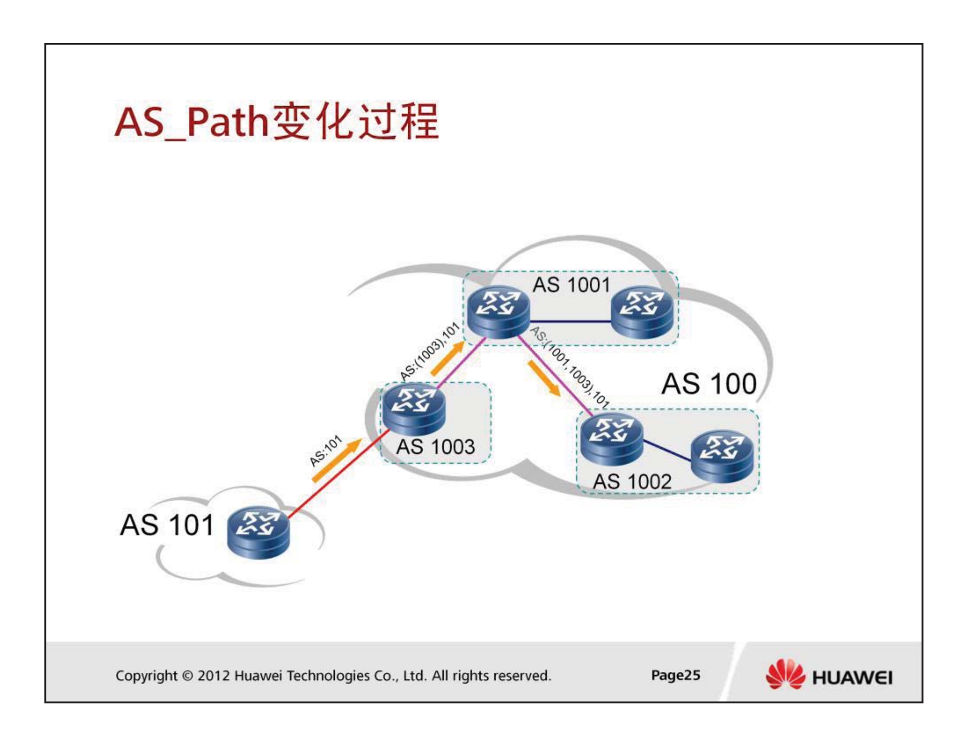
Value Segment Type

1 AS_SET: 由一系列AS号无序地组成，包含在UPDATE消息里。

2 AS_SEQUENCE: 由一系列AS号顺序地组成，包含在UPDATE消息里。

3 AS_CONFED_SEQUENCE: 在本地联盟内由一系列成员AS号按顺序地组成，包含在UPDATE消息中，只能在本地联盟内传递。

4 AS_CONFED_SET: 在本地联盟内由一系列成员AS无序地组成，包含在UPDATE消息中，同样只能在本地联盟内传递。



联盟内采用AS-CONFED来防止子AS间的路由环路。

联盟内的AS-PATH属性变化:

- 联盟内的EBGP会话
  - 子AS号被添加到AS-PATH中的AS-CONFED-SEQUENCE前面
- 联盟内的IBGP会话
  - 不修改AS-PATH
- 外部BGP会话
  - 子AS号从AS-PATH中清除，而大AS号被添加到AS-PATH前面

当一个BGP Speaker通告一条来自其它BGP Speaker的路由时，是否需要修改或者怎样修改AS_PATH属性值，根据对端BGP Speaker的位置所决定:

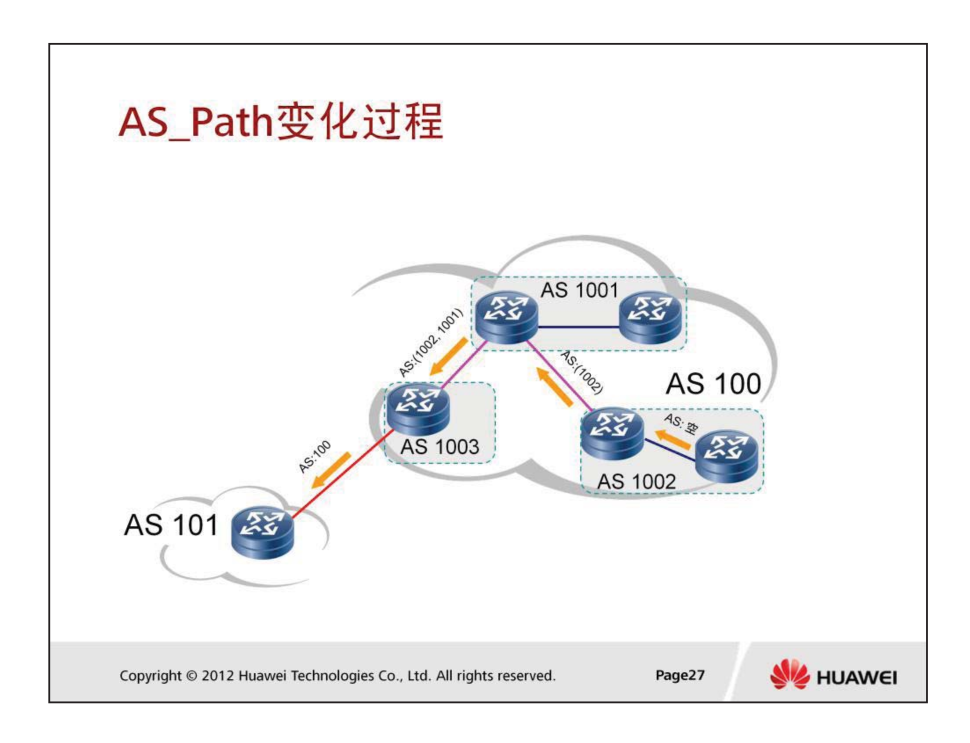
当BGP Speaker和要通告路由去的另外一台BGP Speaker属于同一个AS的时候，通告路由器不需修改AS_PATH。

当BGP Speaker和要通告路由去的另外一台BGP Speaker属于本地联盟内的邻居成员AS的时候，通告路由器需要根据以下情况对AS_PATH进行修改:

- 如果该AS_PATH的类型为AS_CONFED_SEQUENCE，则列前(把AS号放到AS_PATH列表最左的地方)本地AS到该AS_PATH里。
- 如果该AS_PATH的类型不是AS_CONFED_SEQUENCE，则列前本地AS号到新的AS_CONFED_SEQUENCE列表里，并保留原有AS_PATH里的AS号

当BGP Speaker通告路由到另外一台远端BGP Speaker时，并且该远端BGP Speaker属于联盟外的AS时，通告路由器应该按以下条件更新AS_PATH属性：

- 如果发现AS_PATH属性里存在AS_CONFED_SEQUENCE或AS_CONFED_SET时，从AS_PATH属性里删除AS_CONFED_SEQUENCE和AS_CONFED_SET并执行第2步或第3步
- 如果AS_PATH属性里只存在AS_SEQUENCE时，本地系统会加上自身的联盟ID到AS_SEQUENCE里，并作为最后一个经过的AS(放到AS_SEQUENCE里的最左位置)
- 如果删除AS_CONFED_SET/AS_CONFED_SEQUENCE后并没有什么AS_PATH信息，或如果AS_PATH里只存在AS_SET时，更新时应该新添加AS_SEQUENCE到AS_PATH，并加上自己的联盟ID号(本地AS号)

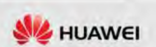


对于外部邻居来说（联盟外的对等体），子AS的拓扑是不可见的。

在更新消息中，已经主动剥去了联盟内已经被修改的AS-PATH属性。

联盟与反射的比较

参考因素	比　　较
多层次	两种方法都支持多层次来进一步增强扩展性。路由反射器支持多级路由反射结构。联盟允许在成员AS内使用路由反射。
策略控制	两者都提供路由选择策略控制，不过联盟可以提供更大的灵活性。
常规IBGP迁移的复杂性	路由反射的迁移复杂性非常低，因为总体网络配置几乎很少发生改变。然而，从IBGP到联盟的迁移需要对配置和网络架构做很大的改变。
能力支持	联盟内的所有路由器必须支持联盟配置能力，因为所有路由器需要支持联盟AS-PATH属性。在路由反射的架构中，只需要反射器支持路由反射能力。然而，在新的分簇设计中，客户也必须支持反射器属性。
IGP扩展	路由反射在AS内需要单一的IGP，而联盟支持单一的或分开的IGP。这可能是联盟比路由反射所具有的最明显的优势。如果IGP达到了其扩展性限制，或者是因为范围太大而难于处理管理任务，那么可以使用联盟来减小IGP路由表的大小。
部署经验	由于更多的服务提供商已经部署了路由反射而非联盟，因此从路由反射中已经获得了更多的经验。
AS合并	实际上AS合并与IBGP扩展性是无关的，但在这里讨论是因为它是联盟的特点之一。一个AS可以和一个已存的联盟合并，这是通过把新的AS作为联盟的一个子AS对待来完成的。

 Page28 



问 题

BGP反射器与联盟主要解决的问题?

请回顾BGP反射器的宣告原则?

BGP联盟的AS-Path变化?

Page29



1.BGP反射器与联盟主要解决的问题?

答：由于BGP通告原则，导致IBGP对等体都必须相互建立IBGP邻居关系，形成IBGP全互连。而IBGP全互连的确可以很好地解决由于BGP通告原则所引起的问题，但同时也带来的另外的一个问题，就是BGP Speaker必须维护更多的IBGP会话数量，因此BGP引入了反射器与联盟。

2.请回顾BGP反射器的宣告原则?

答：当RR收到BGP对等体发来的路由时，使用BGP选路策略来选择最佳路由。在发布学习到的路由信息时，RR按照RFC2796中的规则发布路由。

1) 从非客户机IBGP对等体学到的路由，发布给此RR的所有客户机。

2) 从客户机学到的路由，发布给此RR的所有非客户机和客户机（发起此路由的客户机除外）。

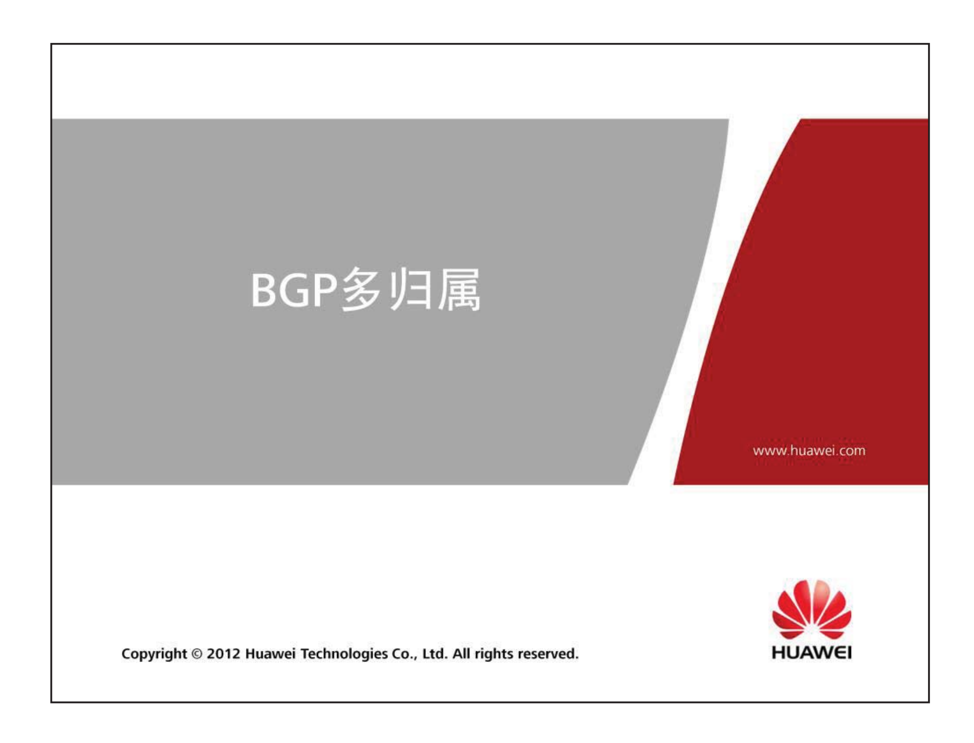
3) 从EBGP对等体学到的路由，发布给所有的非客户机和客户机。

3.BGP联盟的AS-Path变化?

答：由于联盟该技术会让产生更多的成员AS，所以在AS_PATH属性中特别引入两个新类型，为联盟工作，分别是：AS_CONFED_SEQUENCE以及AS_CONFED_SET。

路由信息在成员AS内间传递，每经过一个成员AS后都会加上其AS号在AS_CONFED_SEQUENCE前面，类似于AS_PATH，另外一旦该路由信息发往联盟外的路由器时，联盟的边界路由器会把AS_CONFED_SEQUENCE删除，并把联盟AS号加入到原来的AS_PATH列表里，然后再向外传递。还有一种情况就是从联盟AS外收到一条路由信息，联盟会保留其AS_PATH，并再创建一个AS_CONFED_SEQUENCE，在联盟内使用。


谢谢
www.huawei.com


BGP多归属
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

如今的BGP网络为提供更好更可靠的服务，多归属已经是常见的组网技术。本课程中主要介绍多归属的分类，多归属的好处以及多归属的负载分担。

Page1





培训目标

学完本课程后，您应该能：

- 了解BGP多归属的分类
- 了解BGP多归属的负载分担

Page2




目 录
什么是BGP多归属
多归属负载分担
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page4
HUAWEI

多归属的定义

多于一条外部路径到达本地网络，比如：

- 多路径连接到相同的ISP网络
- 多路径连接到不同的ISP网络

通常会使用两台或两台以上的路由器连接外部网络

- 提供有效的冗余确保可靠的服务
- 实现路由选择和负载分担

 Page5 

只要网络有多于一条路径到达外部，那么它就是多归属的，比如网络通过多条路径到达单个ISP或者多条路径到达不同的ISP。

使用多归属的目的是为了通过冗余链路提供可靠性或实现负载分担。

负载分担即允许路由器将入流量以及出流量分配到多条路径上。多路径可以通过静态路由协议或者动态路由协议进行学习，比如：RIP,OSPF等。

BGP在默认的情况下，只允许使用最佳路径并且不支持负载分担。

本课程将介绍BGP在不同的方案下怎样实现流量的负载分担。

BGP多归属分类

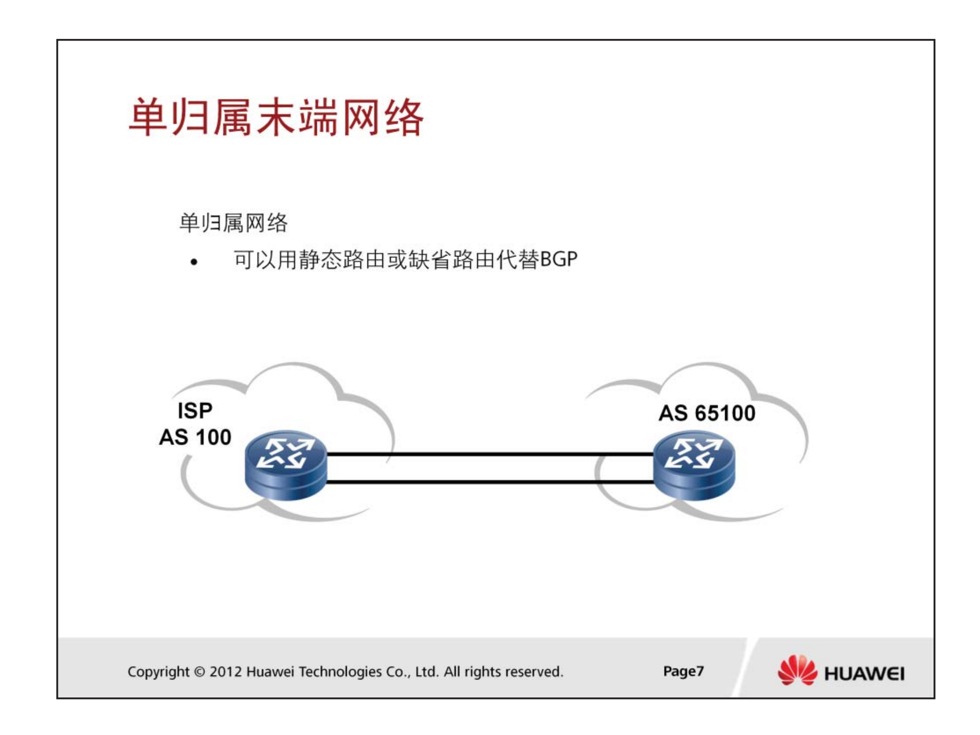
- 单归属末端网络
- 多归属末端网络
  - 单台边界路由器
  - 多台边界路由器
- 多归属到不同ISP网络
  - 单台边界路由器
  - 多台边界路由器

 Page6 

BGP多归属可以分为以下几种类型：

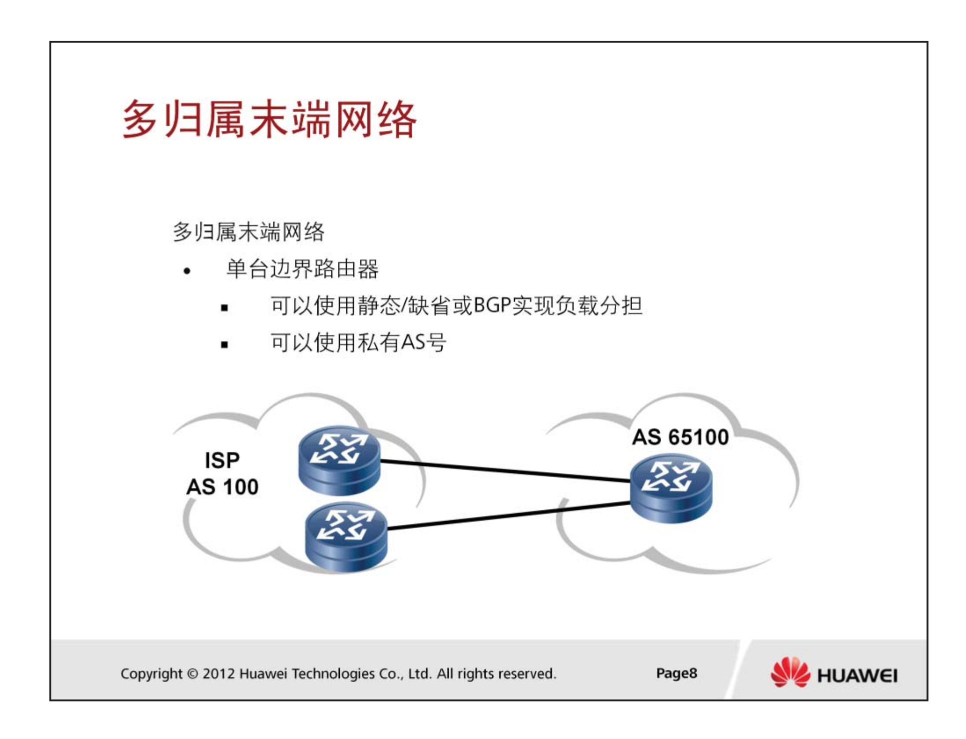
- 单归属末端网络
- 多归属末端网络
  - 单台边界路由器
  - 多台边界路由器
- 多归属到不同ISP网络
  - 单台边界路由器
  - 多台边界路由器

本部分将针对上述几种类型逐一探讨。



单归属末端网络指的是两个AS各通过一台边界路由器互连，两路由器之间可以通过多条链路实现冗余。当前VRP能支持的最大等值路径数为6条。

单归属网络的冗余性不高，适合用在小型的网络里。如图所示，客户AS65100通过两条链路上行到ISP AS100，在这种网络里，客户AS65100可以不需要运行BGP，AS内部可以通过配置一条指向上级设备的静态路由即可。



多归属末端网络的冗余强度就在于单条上行链路失效或者单台ISP路由器失效都不会影响网络正常运行。

在部署这一设计方案时，应该使用BGP来为可能的负载分担提供额外的控制。单个上游提供商的情况使得企业能够使用私有AS号。这意味着企业不需要从注册机构那儿获得唯一公用的、且能被外界看到的AS号。上游提供商可以从接收到的更新信息中清除私有AS号。

在这种设计中使用BGP使企业能够很大程度地影响入流量并更好地控制出流量。在链路带宽不相等的情况下，这特别有用，因为可以使用路由选择策略根据链路带宽来按比例地分担流量。

多归属末端网络 (续)

多归属末端网络

- 多台边界路由器
  - 更好地控制出流量，能够根据按链路带宽按比例地分担流量
  - 边界路由器与边界路由器之间必须运行IBGP会话及核心IGP会话

ISP
AS 100

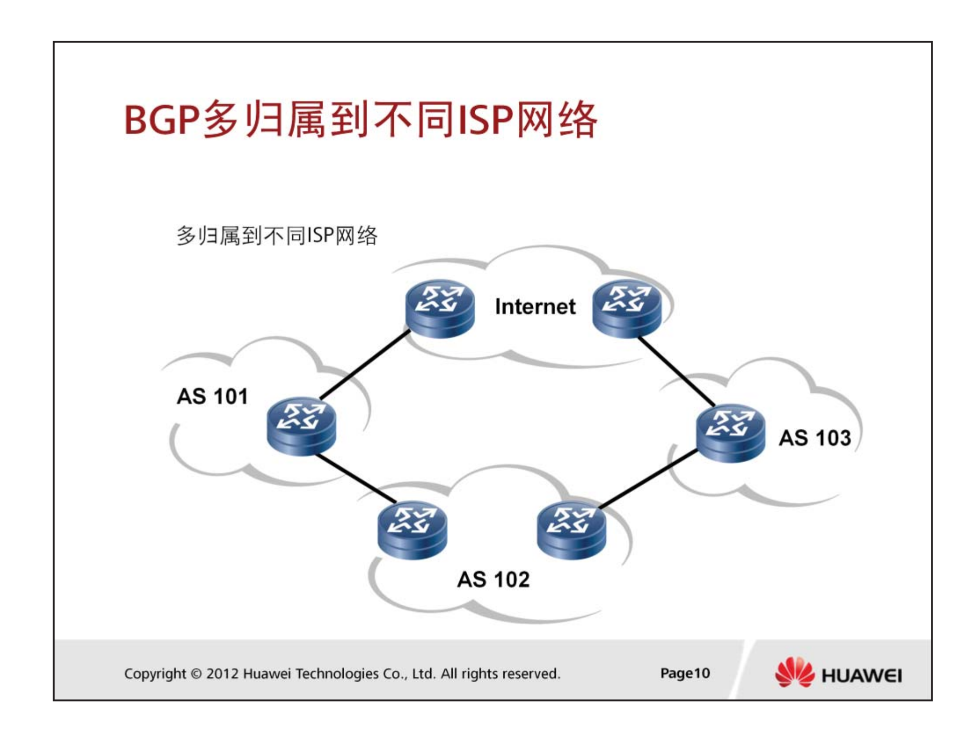
AS 65102

 Page9 HUAWEI

使用单台企业边界路由器可能会产生单点故障。再增加一台或多台边界路由器就消除了最后的故障单点。这种设计要求企业网络中有多台边界路由器，每一台路由器有连接到上游提供商的一条或多条连接。在该设计中，仍然使用单个上游提供商。

在这种设计中，企业仍然可以使用私有AS号。使用私有AS号的主要好处在于企业不需要再花费额外开销去获得一个新的公有AS号。运行BGP的目的在于它能够为企业定义出入境路由选择策略时提供额外的支持。除了需要与上游提供商建立EBGP会话外，企业还应该在边界路由器之间和涉及到有可能为边界路由器提供穿越服务的所有第3层设备之间建立IBGP全连接会话。这一要求保证了流量不会被发送到没有去往目的地址的路由选择信息的设备上。

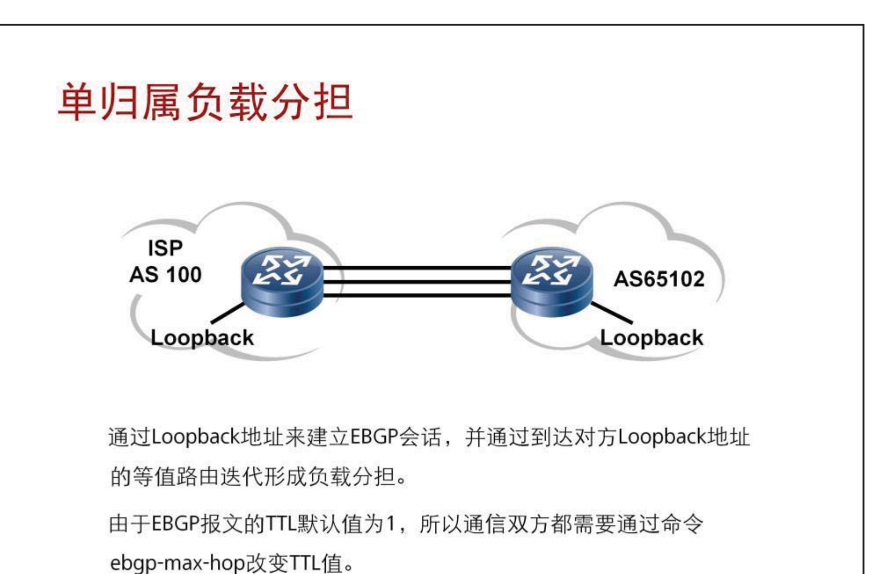
为了防止在边界路由器的上行链路失效的情况下，IGP流量沿着缺省路由流到边界路由器上，边界路由器应该在链路建立并激活的条件下发布缺省路由。这种条件性通告可以基于指向上行接口的静态缺省路由来完成，或者基于从BGP接收到的缺省路由。如果从上游提供商接收到其他路由信息，不要将这些信息引入到边界路由器上运行的IGP进程中，以免引起路由震荡。



在这种设计下，企业边界路由器和它们的上游提供商之间建立了EBGP对等体关系。所有的边界路由器之间IBGP全连接会话。接收到的路由信息量可能只有缺省路由，也可能有完整的路由表。

最常见的负载分担机制只涉及到使用部分的路由选择信息。这可能意味着企业将从上游提供商请求部分路由选择信息并和缺省路由一起使用，或者请求完整的路由表并修改过滤策略，从而获得合理的负载分担。最后使用的方法依赖于企业的实际情况。最简单的方法就是把一条链路用作主用链路，而把其他链路纯粹用作备用链路。最难的就是在多条链路上实现合理的负载分担。

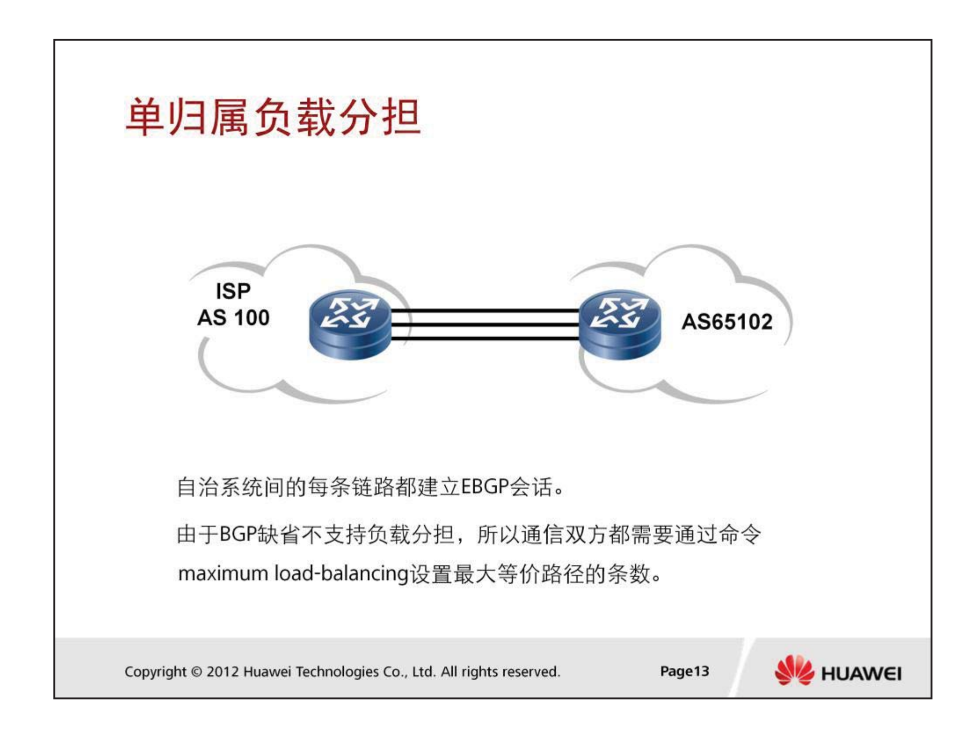

目 录
什么是BGP多归属
多归属负载分担
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page11
HUAWEI



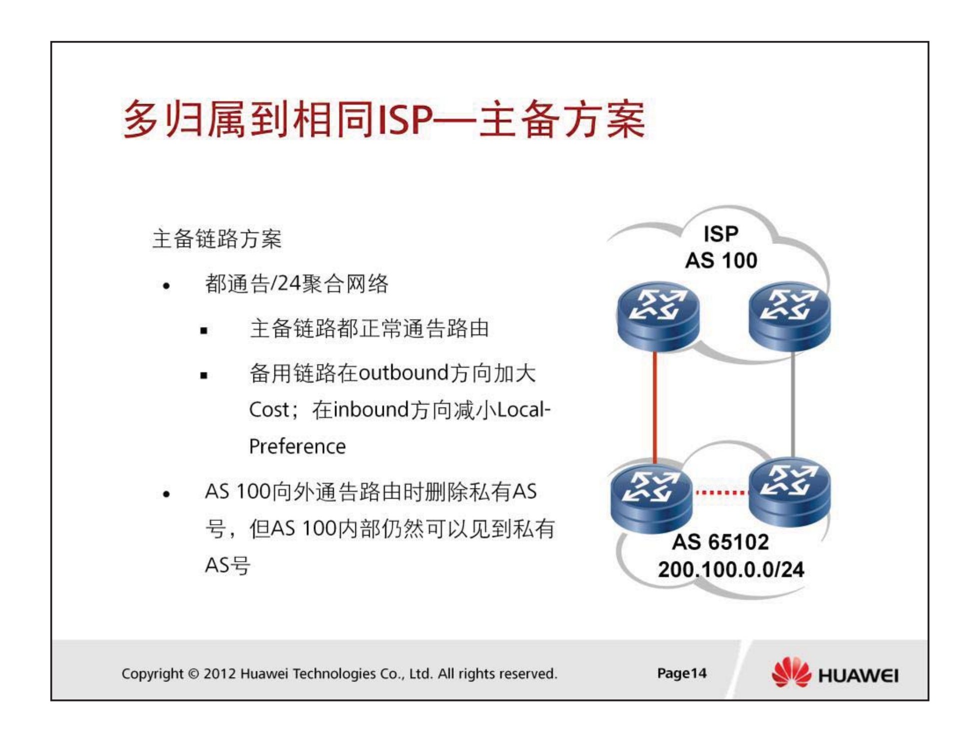


两台路由器之间使用单个EBGP会话， 这个EBGP会话利用Loopback接口作为更新源，代替物理接口。对每一个物理直连接口都配置了一条指向远端Loopback地址的静态路由或者通过IGP来建立其与远端Loopback地址的TCP会话。这种方法解析了下一跳地址，并通过到下一跳的迭代路由来形成负载分担。

必须注意的是需要手动更改EBGP的TTL值大于等于2，否则BGP会话将不能建立。



EBGP多路径提供了另一种在多路径链路上实现负载分担的解决方案。两台路由器之间的每一条链路都被配置了一个EBGP会话。这些EBGP会话被直接捆绑到物理接口地址上。其结果就是两台路由器接收到多条路径信息，每一条链路都有一条路由信息。EBGP多路径允许路由器配置的最大路径数为8条。但是开启BGP多路径特性将导致产生大量的内存需求，通常使用前一种EBGP多跳的解决方案。



本地AS65102通过两链路上行到AS100，其中一链路被配置为主用链路(红色链路)，负责所有流量的转发；另一链路被配置为备用链路(灰色链路)，当主链路失效时，所有流量都转向备用链路上。

如图所示，主备链路都正常通告200.100.0.0/24；但是备用链路通过路由策略改变MED值，另外也通过路由策略减小在Inbound方向上接收到的所有前缀的Local_Pref值。通过这样的配置来实现出流量和入流量都分布在主用链路上。

私有AS号可以在本地AS内使用。当AS100向外通告路由时，私有AS信息会被删除。

多归属到相同ISP—负载分担方案

入流量负载分担
- 两链路都通告/24路由
- 分割/24路由为两条/25路由，每条链路上通告一条
- 可以通过再分割直到实现等值的负载分担

出流量负载分担
- 使用缺省路由
- 接收上游的路由信息
  - 通过“最近的出口”实现基于出流量的负载分担

ISP
AS 100
AS 65102
200.100.0.0/24

Page15
HUAWEI

该方案通过多台本地路由器来提供多条链路连接到ISP。其中两台路由器分别通过EBGP对等体连接到上级AS100，实现本地网络200.100.0.0/24的出入流量的负载分担。

本地路由器正常通告200.100.0.0/24位路由，同时将路由细分为两条/25位的前缀，再通告给上级AS100，但这里需要留意，由于/25位的前缀很有可能并不在设备的IP路由表里，所以需要分别在两台本地路由器上加上静态路由：ip route-static 200.100.0.0 25 null 0 以便/25位前缀能成功向外通告。

细分前缀的目的是为了让上级设备尽可能匹配其路由，实现基于入流量的负载分担。如上图所示，如/25位前缀还不能实现理想的负载分担的话，可以再进一步的细分，直到接近等值的负载分担。

怎样实现出流量的负载分担？ 我们可以通过配置路由策略只接收部分感兴趣的路由信息，针对不同目的地址配置路由策略来实现负载分担。如：到达100.100.1.0/24的流量通过左边链路上行，而100.100.2.0/24的流量从右边链路上行。

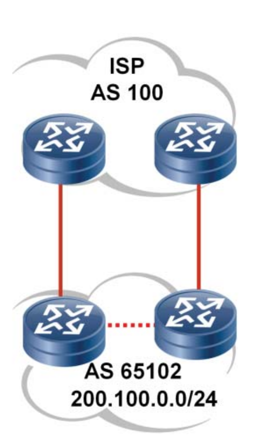
实际上，平衡流量最简单的办法就是使用缺省路由，这可以提供均衡的

流量；不过，它产生次优路径的可能性非常大。如果企业多归属到同一个上游ISP，那么使用缺省路由很可能就是最简单的解决方案。

多归属到相同ISP—负载分担方案 (续)

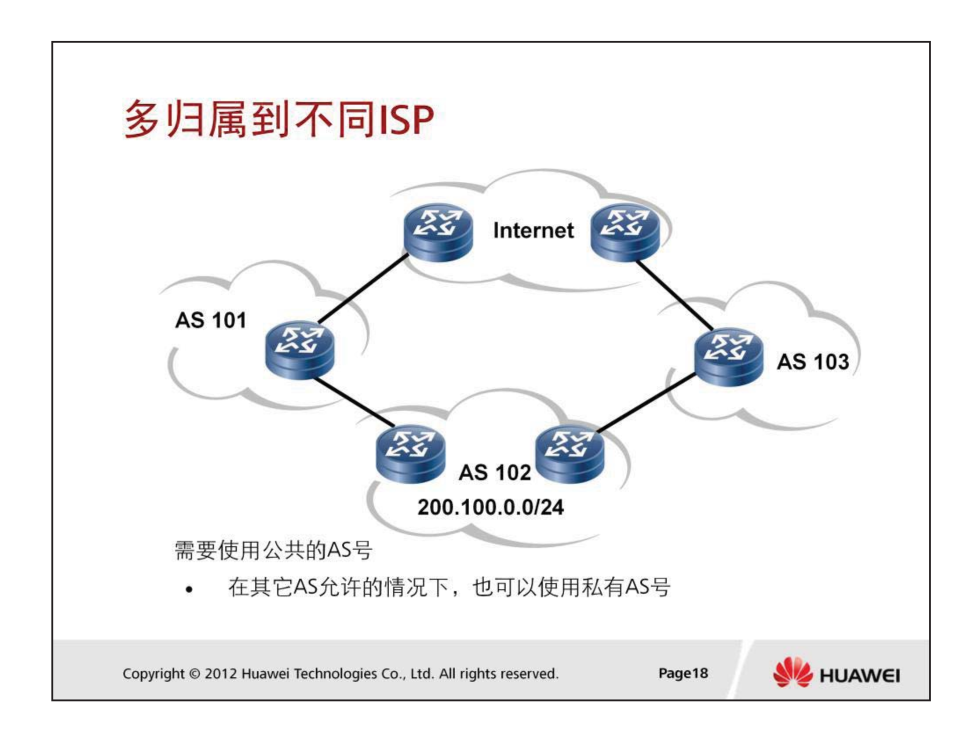
负载分担方案（续）

- AS 100的边界路由器只接收下游的路由信息
- AS 100删除AS 65102里的子网信息
- AS 100向外通告路由时需要删除私有AS号





Page17 HUAWEI



在这种设计下，企业边界路由器和它们的上游提供商建立了EBGP关系。所有的边界路由器之间和可能为企业边界路由器提供穿越服务的任何其他的第3层设备之间也有IBGP全连接。接收到的路由信息量可能只有缺省路由，也可能是到完整的路由表。这种情况下，接收路由信息的状况和使用单台路由器的状况是相同的。

最常见的负载分担机制只涉及到使用部分的路由选择信息。这可能意味着企业将从上游提供商请求部分路由选择信息并和缺省路由一起使用，或者请求完整的路由表并修改过滤策略，从而获得合理的负载分担。最后使用的方法依赖于企业的实际情况。最简单的方法就是把一条链路用作主用链路，而把其他链路纯粹用作备用链路。最难的就是在多条链路上实现合理的负载分担。

多归属到不同ISP —主备方案

主备链路方案

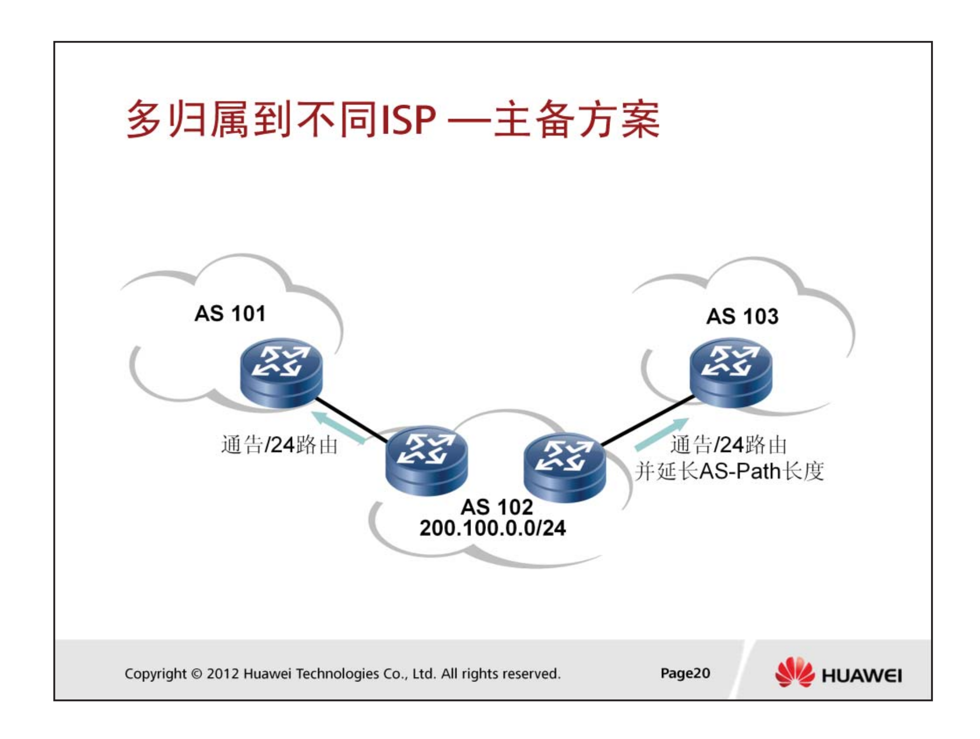
- 两链路上都通告/24聚合路由
  - 主用链路采用标准路由通告
  - 备用链路延长AS-Path长度通告路由

当一链路失效时，通过另一条链路保证连接



该方案与多归属到相同ISP的实现方法相似。

在决定使用哪条链路的时候，根据AS_PATH属性来选择更优的路径进入企业网络。



如图所示，AS102通过两条上行链路都通告200.100.0.0/24路由，主用链路采用标准形式通告，备用链路通过增加AS-Path长度来通告，流量根据AS-Path属性将优选主用链路进入企业网络，同时，备用链路提供冗余。

多归属到不同ISP —负载分担方案

负载分担方案

- 两链路上都正常通告/24聚合路由
- 同时把/24聚合路由细分为两/25路由，每链路通告一条
  - 实现入流量的负载分担
- 修改通告路由的AS_PATH长度

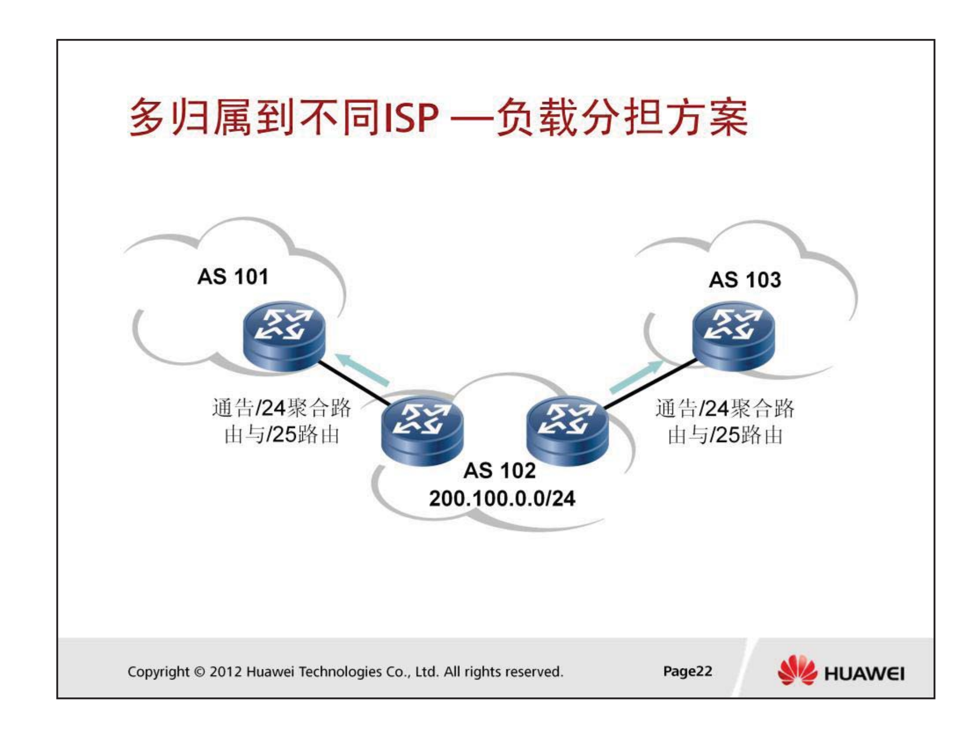
当一链路失效时，通过另一条链路保证连接

 Page21



该方案与多归属到相同ISP的实现方法相似，细分前缀的目的是为了让上级设备尽可能匹配其路由，实现基于入流量的负载分担。

但是通告更详细的前缀的方法并非总是可用的，在这种情况下，可以通过修改路由的AS_PATH长度来解决。当把路由通告给不同的ISP后，监控链路的利用率，如果流量总是选择其中一条链路，那么就在高利用率的路径上增加AS_PATH长度，然后继续监控链路的利用率。



如图所示， AS102通过两条上行链路都通告200.100.0.0/24路由，同时，把/24聚合路由细分为两条/25路由，每条链路通告一条以实现负载分担，如果效果不理想，则可以继续通过修改AS_PATH的长度的方法来实现，在高利用率的路径上增加AS_PATH长度，然后继续监控链路的利用率。



问 题

我们为什么需要多归属的网络?

多归属到不同的ISP，怎样实现入流量的负载分担?

 Page23



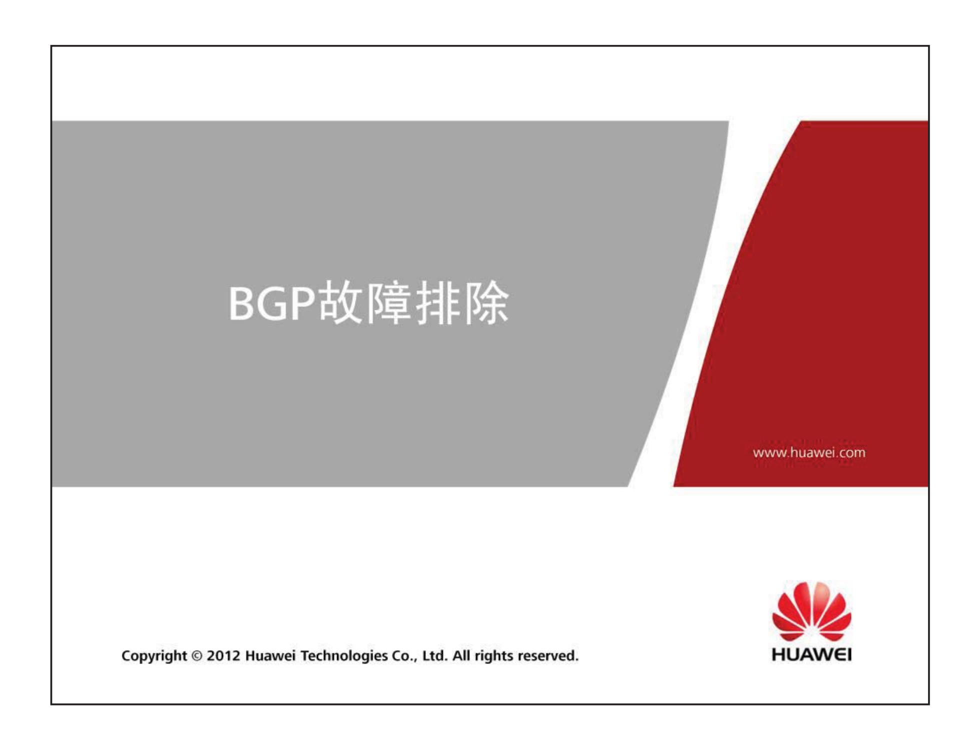
Q: 我们为什么需要多归属的网络?

A; 多归属网络为我们提供更好的冗余性，更多的流量控制方案。

Q: 多归属到不同的ISP，怎样实现入流量负载分担?

A: 对于多归属到不同的ISP，入流量的负载分担可以通过增加AS_PATH的长度来实现，但需要注意，AS_PATH的长度建议每次只增加一个，另外每增加一次都可监控链路流量，才决定是否需要再一次增加AS_PATH的长度。


谢谢
www.huawei.com


BGP故障排除
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

BGP作为一个复杂的域间路由协议，经常出现各种各样的故障，如何去定位故障的原因以及准确的排除故障，都需要建立在对协议运作非常了解的基础上。本课程就是通过实例来加深对BGP的理解，提高故障处理的能力。

Page1





培训目标

学完本课程后，您应该能：

- 知道故障排除的基本步骤
- 了解BGP故障的基本排错方法

Page2





目 录

Page3





目 录

BGP故障处理流程

BGP对等体建立故障

BGP路由学习故障

BGP路径选择故障

Page4



BGP故障原因

产生BGP故障的原因主要分为以下三个部分：

- 错误的配置
  - 对BGP不了解或者错误的配置脚本都会导致配置错误，从而使BGP Speaker产生一系列不明确的错误
- 人为导致
  - 其实大多数情况下都由于人为错误导致BGP产生故障，人为错误主要有：使用错误的命令，组网设计存在缺陷等
- 版本问题
  - 不了解设备版本信息导致的故障，如：某版本支持但另外一个版本不支持等

 Page5



产生BGP故障的原因主要分为以下三点：

- 错误的配置
  - 对BGP了解不够以及错误的配置脚本都会导致错误配置，从而使BGP Speaker产生一系列不明确的错误
- 人为导致
  - 其实不管BGP还是其它一些路由协议，大多数情况下都由于人为错误导致，人为错误主要有：使用错误的命令、组网设计存在缺陷等
- 版本问题
  - 不了解设备版本信息导致的故障，如：某版本支持但另外一个版本不支持等

BGP故障处理程序

发现故障

- 记录故障现象

收集信息

- 通过BGP查看命令收集信息

处理故障

- 整理收集的故障现象并根据以往的经验制定Checklist，严格按照Checklist的步骤逐步进行尝试，直到故障排除

经验总结

- 对已经排除的故障，把故障现象以及解决方法都记录下来

Page6



故障排除步骤:

1.发现故障

- 记录故障现象

2.收集信息

- 通过BGP查看命令收集信息

3.处理故障

- 整理收集的故障现象并根据以往的经验制定Checklist，严格按照Checklist的步骤逐步进行尝试，直到故障排除。如果最后还是没法解决故障，请联系技术支持人员。

4.经验总结

- 对已经排除的故障，把故障现象以及解决方法都一一记录。目的是为后续的工程师提供有用的经验。



目 录

Page7



BGP建立邻居关系

TCP连接

- BGP连接建立在TCP会话基础之上，端口号为179

IP连通性

- BGP对等体在大多数的情况下，需要静态路由或IGP提供可达性

OPEN消息的交互

- OPEN消息是BGP建立对等体关系时交互信息的主要报文，交互的信息主要有：AS号、更新源地址以及其他参数

 Page8



以下提到的三点，任何一点出现错误，都会导致BGP无法建立邻居关系：

TCP连接

- BGP连接建立在TCP会话基础之上，端口号为179。也就是说一旦端口179被禁用，BGP则无法建立邻居关系。

IP连通性

- BGP对等体在大多数的情况下，需要静态路由或IGP提供可达性。

OPEN消息的交互

- OPEN消息是BGP建立对等体关系时交互信息的主要报文，交互的信息主要有：AS号、更新报文的源地址以及其他参数。

BGP建立邻居关系 (续)

EBGP多跳

- 建立EBGP邻居关系时，报文的默认TTL值为1，当建立非直连邻居关系时需要手动修改TTL值

其它问题

- 物理连接的不稳定导致振荡 (经常UP/DOWN)

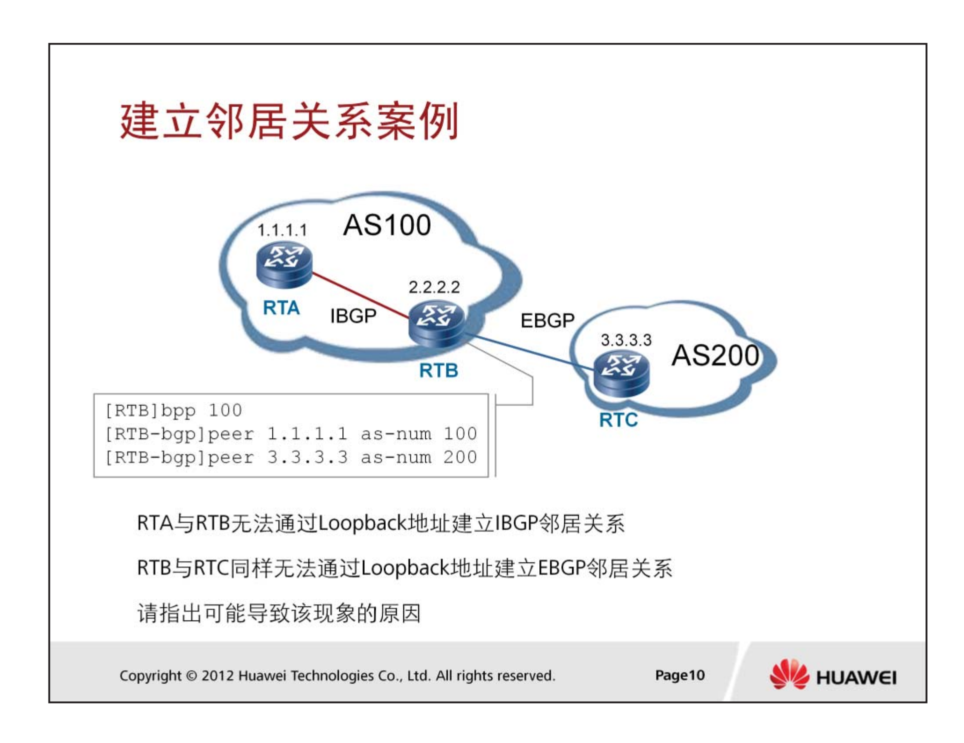
 Page9



除了TCP连接，IP连通性及OPEN消息需要注意以外，错误配置BGP也是一个导致故障的常见原因。

EBGP在默认的情况下更新报文的TTL值为1，所以当BGP通过Loopback地址或非直连接口建立邻居关系的时候，就需要通过命令修改其TTL的值。另外，除EBGP邻居关系的建立需要注意以外，IBGP也有需要注意的地方，比如说：建立邻居关系的源端口不能相互匹配也会导致IBGP邻居关系无法建立。

接下来就是其它问题了，主要体现在物理连接上面。由于物理连接的不稳定，导致端口的振荡也是经常出现的故障。



如图所示，AS100内的RTB希望与RTA建立IBGP邻居关系，希望与AS200的RTC建立EBGP邻居关系。

现在故障现象是RTB与RTA无法建立IBGP邻居关系，与RTC也无法建立EBGP邻居关系。

根据之前的分析，邻居关系无法建立主要有以下原因：

TCP 179端口被禁用

没有IP连通性

OPEN消息参数不正常

EBGP/IBGP配置有误

物理层以及其它故障



首先在RTB上通过命令“display bgp peer”查看BGP对等体信息，留意两BGP对等体状态都是Active，则表明没有建立TCP连接。

收集信息—TCP信息

可以检查本地TCP端口

- 本地端口179已经打开，并处于“Listening”状态，表明本地并没有禁用179端口

```
[RTB]display tcp status
TCPCB      Local Add:port        Foreign Add:port      State
048b1f64   0.0.0.0:23            0.0.0.0:0             Listening
04d18724   0.0.0.0:179           1.1.1.1:0             Listening
04d2fc84   0.0.0.0:179           3.3.3.3:0             Listening
04d30224   10.1.1.2:49554        3.3.3.3:179           Syn_Sent
```

 Page12 

用命令 “display tcp status” 查看本地路由器的TCP端口状态。

可以看出有两条目的179端口正处于“Listening”状态，表明本地TCP端口179没有被禁用。

收集信息—TCP信息 (续)

打开调试信息，查看TCP报文交互情况

- 从下图可以判断：
  - RTA和RTC都主动发起端口号为179的TCP连接，表明RTA与RTC并没有禁用TCP179端口

```
<RTB>debug tcp packet
*0.4064578 RTB SOCKET/8/TCP PACKET:
1154168666: Input: task = ROUT(9), socketid = 7, state = Syn_Sent,
src = 1.1.1.1:179, dst = 10.1.1.2:50842,
seq = 0, ack = 521746511, flag = ACK RST,
window = 0
*0.4093578 RTB SOCKET/8/TCP PACKET:
1154168695: Input: task = ROUT(9), socketid = 5, state = Syn_Sent,
src = 3.3.3.3:179, dst = 10.2.2.1:54472,
seq = 0, ack = 525522511, flag = ACK RST,
window = 0
```

 Page13



怎样通过命令查看远端对等体的TCP端口是否被禁用？

我们可以通过debug命令，查看TCP的debug信息。如图所示，收到两个TCP的报文，分别来自1.1.1.1和3.3.3.3，并且端口都是179。这就可以证明远端的TCP端口179也没有被禁用。

另外我们再细心看看，源地址为1.1.1.1，而目的地址是10.1.1.2，这说明了RTA通过Loopback地址与RTB的物理接口地址建立邻居关系，同样RTC也是通过Loopback地址与RTB的物理接口地址建立邻居关系。

收集信息—IP连通性

通过PING命令查看IP连通性(注：加上“-a”参数，指定ping的源地址)

- 从RTB到RTA没有IP连通性问题

```
[RTB]ping -a 2.2.2.2 1.1.1.1
  PING 1.1.1.1: 56  data bytes, press CTRL_C to break
    Reply from 1.1.1.1: bytes=56 Sequence=1 ttl=255 time=32 ms
    Reply from 1.1.1.1: bytes=56 Sequence=2 ttl=255 time=32 ms
    Reply from 1.1.1.1: bytes=56 Sequence=3 ttl=255 time=32 ms
    Reply from 1.1.1.1: bytes=56 Sequence=4 ttl=255 time=32 ms
    Reply from 1.1.1.1: bytes=56 Sequence=5 ttl=255 time=32 ms

  --- 1.1.1.1 ping statistics ---
    5 packet(s) transmitted
    5 packet(s) received
    0.00% packet loss
    round-trip min/avg/max = 32/32/32 ms
```

 Page14



TCP端口被禁用这一可能性已经被排除。那么接下来的就是查看两BGP Speaker之间是否有IP连通性。而检查IP连通性最常用到的就是“ping”命令。

为了更准确的检查两端口是否可达，我们可以在“ping”命令后面加上“-a”参数，指定源地址。

结果如上所示，从源地址2.2.2.2到目的地址1.1.1.1，并没有任何IP连通性的问题。

收集信息—IP连通性 (续)

通过PING命令查看IP连通性(注：加上 “-a”参数，指定ping的源地址)

- 从RTB到RTC存在IP连通性问题

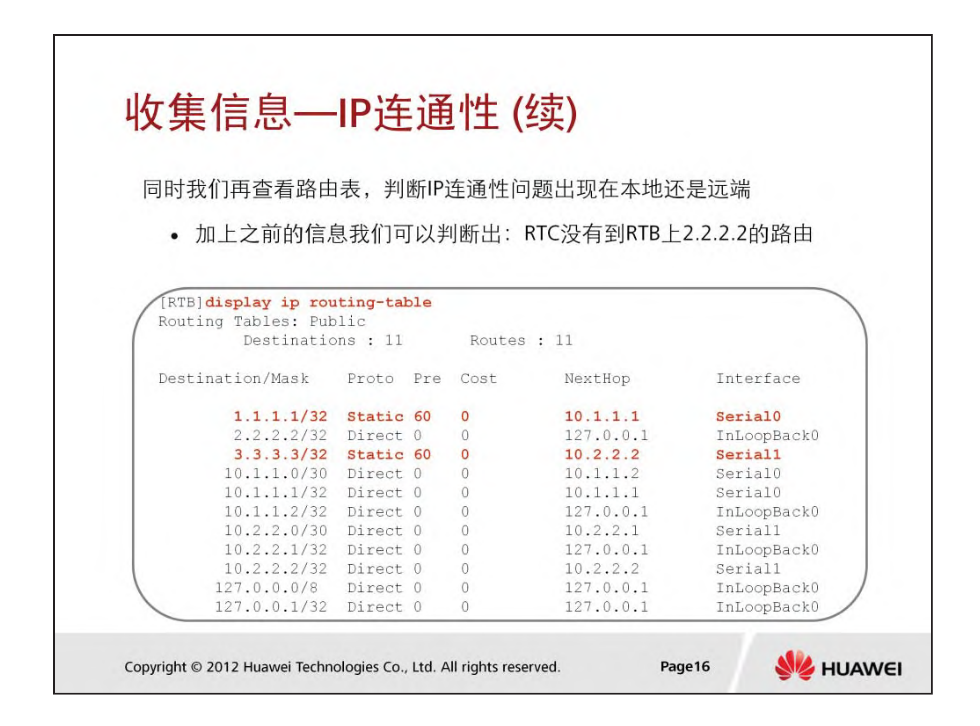
```
[RTB]ping -a 2.2.2.2 3.3.3.3
  PING 3.3.3.3: 56  data bytes, press CTRL_C to break
    Request time out
    Request time out
    Request time out
    Request time out
    Request time out

  --- 3.3.3.3 ping statistics ---
    5 packet(s) transmitted
    0 packet(s) received
    100.00% packet loss
```

 Page15



但RTB与RTC之间则存在IP连通性的问题。RTB的Loopback地址无法PING通RTC的Loopback地址。



查看RTB上的IP路由表，可以看出两条静态路由，目的地址分别为：1.1.1.1和3.3.3.3。这说明问题并不是在RTB路由器上，而是RTC并没有到达RTB的回程路由。

收集信息—配置信息

收集指定路由器的相关配置命令

- 没有修改EBGP的TTL值，导致RTB与RTC之间的邻居关系无法建立
- RTC上指定2.2.2.2的AS号不匹配

```
[RTB]display current-configuration configuration bgp
#
bgp 100
 peer 1.1.1.1 as-number 100
 peer 3.3.3.3 as-number 200
```

```
[RTC]display current-configuration configuration bgp
#
bgp 200
 peer 2.2.2.2 as-number 201
```

 Page17



接下来就是查看RTB以及RTC的配置脚本。

没有修改EBGP的TTL值，导致RTB与RTC之间的邻居关系无法建立；

RTC上指定2.2.2.2的AS号不匹配；

制定Checklist

根据收集的信息以及经验，总结出故障处理的Checklist

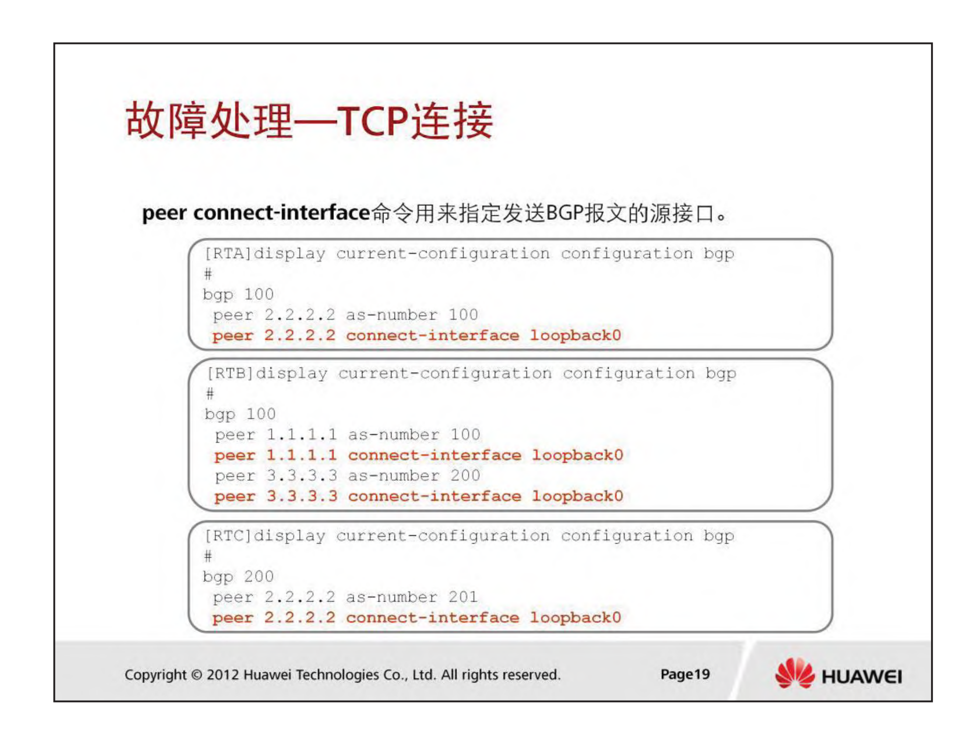
- TCP连接
  - BGP邻居更新的源地址不匹配
- IP连通性
  - RTC没有到达RTB的路由
- 配置信息
  - 修改RTB以及RTC上EBGP邻居更新信息的TTL值
  - BGP配置的AS号不匹配

Page18



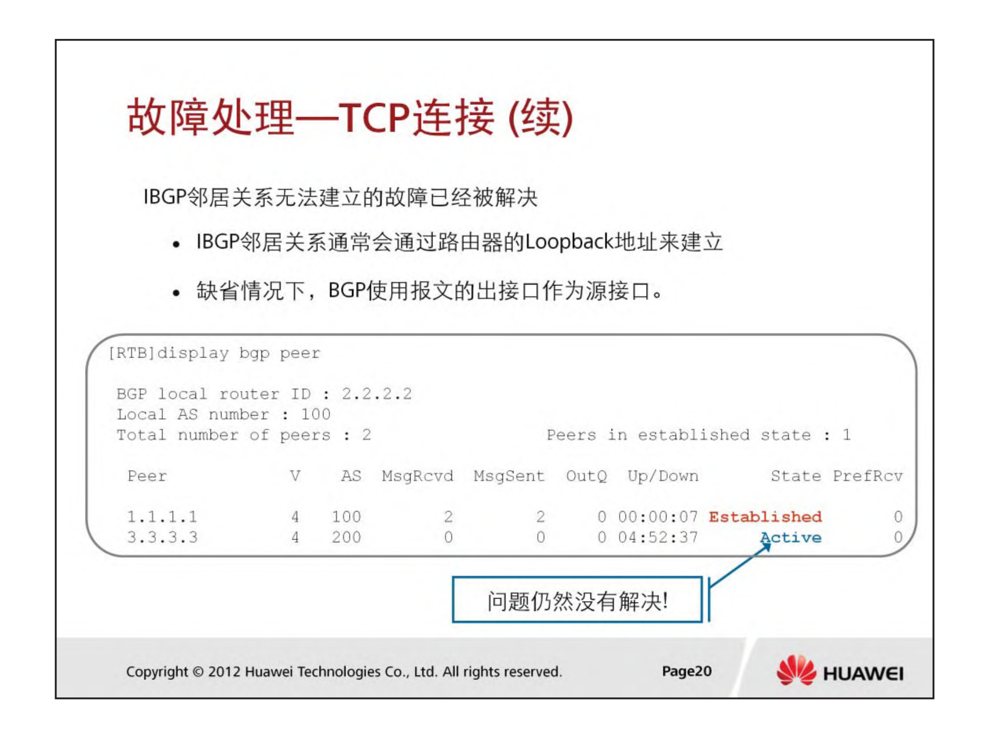
根据之前所收集的信息，我们就可以制定Checklist。

- TCP连接
  - BGP邻居更新的源地址不匹配
- IP连通性
  - RTC没有到达RTB的路由
- 配置信息
  - 修改RTB以及RTC上EBGP邻居更新信息的TTL值
  - BGP配置的AS号不匹配

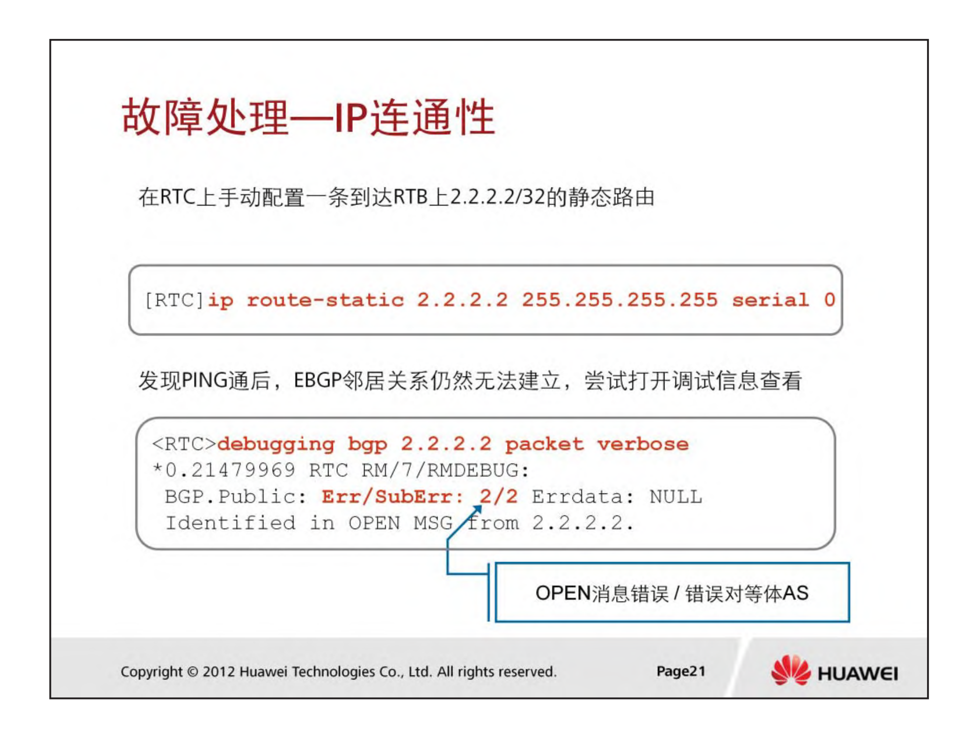


首先先解决TCP连接的问题。

IBGP对等体建立邻居关系默认是通过最佳源接口地址来建立，但我们现在需要的是通过Loopback地址来建立邻居关系，所以必须通过命令“peer connect-interface”来改变发送BGP报文的源接口。



修改配置后，IBGP的邻居关系已经成功建立，但EBGP的邻居关系还是处于Active状态。



Checklist第二步，在RTC上增加一条到2.2.2.2的静态路由，解决IP连通性问题。但EBGP对等体还没法建立，通过“debug”命令看出，EBGP邻居关系的AS号不匹配。

故障处理—配置信息

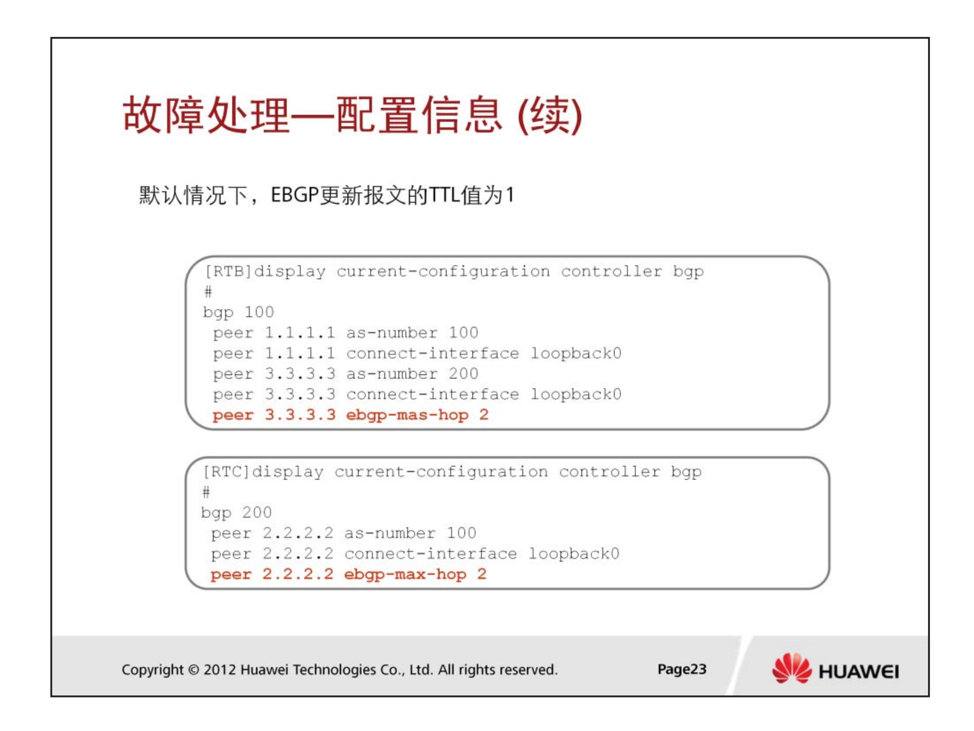
针对OPEN消息的错误进行配置

```
[RTB]display current-configuration configuration bgp
#
bgp 100
 peer 1.1.1.1 as-number 100
 peer 1.1.1.1 connect-interface loopback0
 peer 3.3.3.3 as-number 200
 peer 3.3.3.3 connect-interface loopback0
```

```
[RTC]display current-configuration configuration bgp
#
bgp 200
 peer 2.2.2.2 as-number 100
 peer 2.2.2.2 connect-interface loopback0
```

 Page22 HUAWEI

在RTC上修改后，故障仍然存在。



再增加一条命令 “peer ebgp-max-hop”修改EBGP更新报文的默认TTL值为2，故障排除。



根据之前的故障排除例子，我们可以总结出：

公共注意事项：

- TCP端口179是否被禁用
- IP连通性

IBGP邻居关系建立的注意事项：

- 指定更新源地址

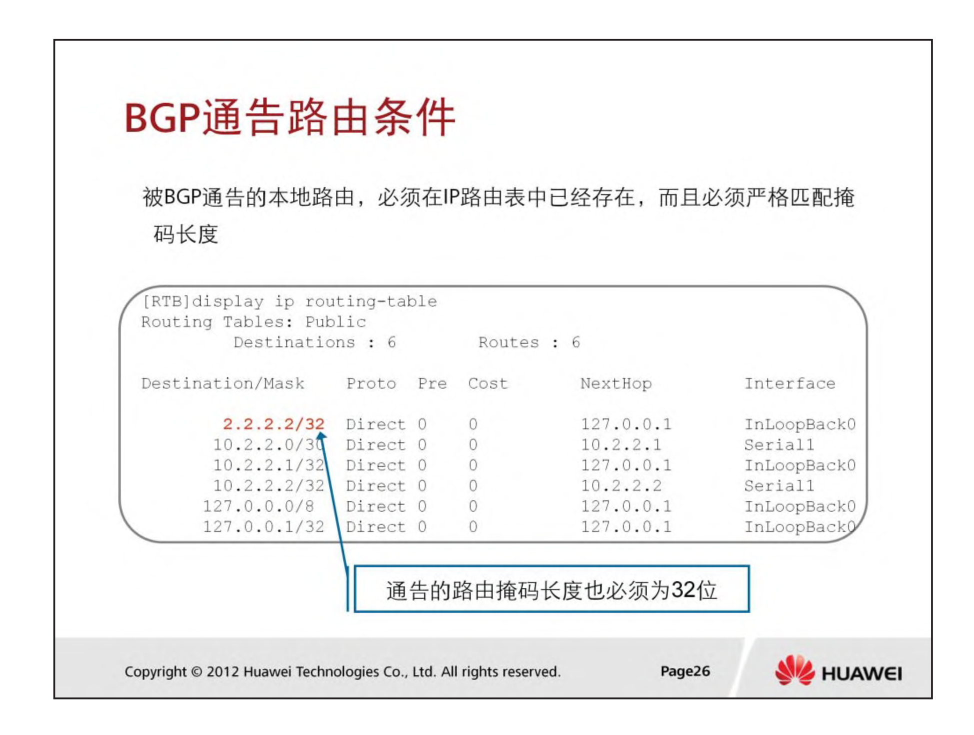
EBGP邻居关系建立的注意事项：

- EBGP多跳问题



目 录

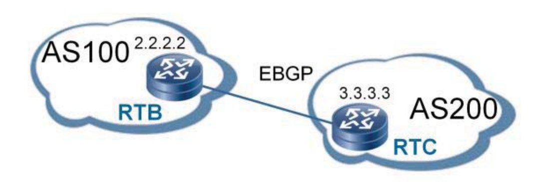




BGP只通告存在于IP路由表中的前缀，并且必须严格匹配其掩码长度。

如图所示，BGP通告2.2.2.2/32这条路由信息的时候必须带上32位长度的掩码，否则BGP会按照IP地址的分类加上自然掩码。

BGP通告路由条件—举例



RTB与RTC的EBGP邻居关系已经成功建立，但RTC上无法学习到RTB的路由信息：2.2.2.2/32。

 Page27



故障现象：RTB与RTC已经成功建立邻居关系，但RTC上无法学习来自RTB的主机路由：2.2.2.2/32。

BGP通告路由条件—举例 (续)

```
[RTB]display current-configuration configuration bgp
#
bgp 100
 peer 10.2.2.2 as-number 200
 #
 ipv4-family unicast
  network 2.0.0.0
  undo synchronization
  peer 10.2.2.2 enable
```

如果通告路由的时候没有带上掩码长度，BGP则会自动按照其IP地址的类别加上自然掩码

```
network 2.2.2.2 255.255.255.255
```

 Page28 

通过“display cu configuration bgp”查看BGP配置，可以看出在BGP 100的进程里已经通告了路由信息2.0.0.0，但由于BGP会自动加上自然掩码的长度，也就是8位掩码，与IP路由表里的32位掩码不匹配。所以BGP无法成功通告这一条路由信息。

通过带上掩码长度的通告，“network 2.2.2.2 255.255.255.255”问题得到解决。

成为BGP路由的方法

通过network命令

- 被network命令通告的前缀必须存在于IP路由表里

通过aggregate命令

- 被aggregate命令通告的前缀必须存在于BGP路由表里

通过import命令

- 被import命令通告的前缀也必须是存在于IP路由表里

Page29



关于怎样才能成为BGP的路由，总共有三种方法：

通过network命令

- 其通告的前缀必须存在于IP路由表里，并且需要严格匹配IP路由表里的掩码长度。如果通告的前缀并不存在于IP路由表中，则可以通过添加一条下一跳为空的静态路由解决。

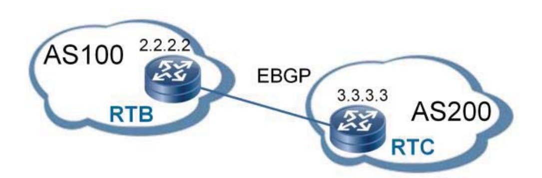
通过aggregate命令

- 其通告的前缀必须存在于BGP路由表里。

通过import命令

其通告的前缀必须存在于IP路由表里。

成为BGP路由的方法—举例



分别通过network, aggregate以及import命令把路由注入到BGP路由表中

RTB通告2.2.0.0/16网段路由给RTC，但RTC没有收到该网段的路由信息

 Page30 

我们仍然以之前的例子来说明，两路由器RTB与RTC。现在RTB并不想通告主机路由，即：2.2.2.2/32，而是想只通告16位掩码的路由，问题应该怎样解决?

BGP通告路由条件—举例 (续)

可以通过配置一条下一跳为空的静态路由，成功通告16位掩码的路由前缀

```
ip route-static 2.2.0.0 16 null 0
```

network命令也能用于聚合路由，但需要在IP路由表里添加相关路由信息

 Page31



配置一条静态路由：ip route-static 2.2.0.0 16 null 0 主要目的是为了在IP路由表中添加一条16位掩码长度的路由信息。

这样RTB就能成功通告16位掩码长度的路由了。

另外这里再说明一下， “network”命令结合静态路由，其实也可以完成聚合路由的工作，比如以上的例子。虽然能完成聚合的工作，但配置繁琐，功能也不如 “aggregate”命令齐全。

BGP通告路由回顾

BGP邻居关系建立以后，通过UPDATE消息交换路由信息

BGP只通告最佳路由给对等体

BGP对等体收到来自EBGP的更新，通告给所有其它对等体

BGP对等体收到来自IBGP的更新，只通告给所有EBGP对等体（确保同步的情况下）

 Page32

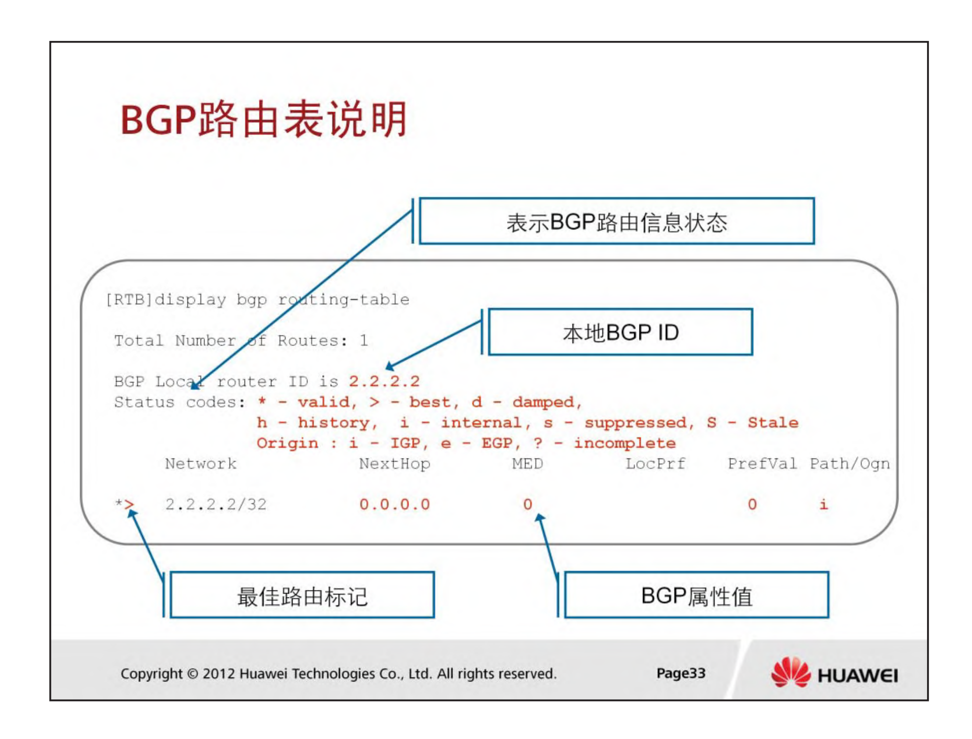


BGP邻居关系建立以后，通过UPDATE消息交换路由信息

BGP只通告最佳路由给对等体

BGP对等体收到来自EBGP的更新，则通告给所有其它对等体

BGP对等体收到来自IBGP的更新，只通告给所有EBGP对等体（确保同步的情况下）



BGP路由表的学习。

主要关心的是BGP Route-id、状态代码、路由条目以及属性。

BGP通告路由注意事项

BGP通告路由的时候需要注意:

- Network通告的路由必须存在于IP路由表中
- Aggregate通告的聚合路由必须存在于BGP路由表中
- Import引入路由通常结合路由策略一起使用

Page34





Page35



BGP路径选择回顾

1.如果此路由的下一跳不可达，忽略此路由

2.Preferred-Value数值高的优先

3.Local-Preference值最高的路由优先

4.聚合路由优先于非聚合路由

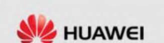
5.本地手动聚合路由的优先级高于本地自动聚合的路由

6.本地通过**network**命令引入的路由的优先级高于本地通过**import-route**命令引入的路由

7.AS路径长度最短的优先

8.比较Origin属性，IGP优于EGP，EGP优于Incomplete

9.选择MED较小的路由

 Page36 

BGP路径选择回顾 (续)

10.EBGP路由优于IBGP路由

11.BGP优先选择到BGP下一跳的IGP度量最低的路径

- 当以上全部相同，则为等价路由，可以负载分担
  - 注：AS-Path必须一致
  - 当负载分担时，以下3条原则无效

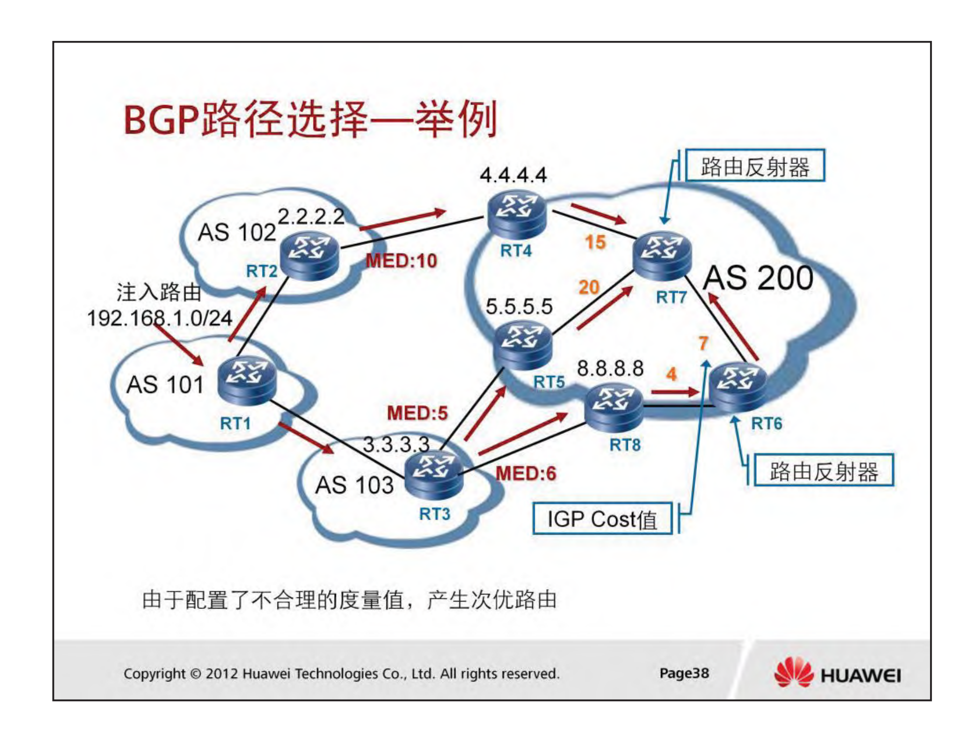
12.比较Cluster List长度，短者优先

13.比较Originator_ID(如果没有Originator_ID，则用Router ID比较)，选择数值较小的路径

14.比较对等体的IP地址，选择IP地址数值最小的路径

　Page37





由于配置了不合理的度量值，产生次优路由。其中可以影响BGP选路的度量值主要有：AS_PATH、MED和IGP Cost值。

RT7收到来自RT4、RT5以及RT6的三条路由更新，分别都带有各自的度量值(如图所示)，并且RT4、RT5和RT8上都已经配置“next-hop-local”。

路径选择过程(情况一)

假如RT7先收到来自RT5或者RT6的路由更新，然后再收到RT4的路由更新

RT7的路由表

路径	BGP下一跳	AS-PATH	MED	IGP度量
1	RT5	103 101	5	20
2	RT8	103 101	6	11
3	RT4	102 101	10	15

 Page39



BGP会根据路由信息接收的先后顺序，一个一个地比较，最终被选举的路径将被BGP认为是最佳路径。

根据BGP的路径选路原则，首先通过比较路径1和路径2这两条路由，由于具有相同的AS_PATH，MED值越低越好，所以路径1被认为比路径2要好；然后再通过比较路径1与路径3，在默认的情况下，AS_PATH不相同则不会比较MED值，所以比较IGP的度量值，最终路径3会被BGP认为是最佳路由。

但我们可以通过组网图看出，路径3并不是最佳的路由。

路径选择过程(情况二)

假如RT7先收到来自RT4的路由更新，然后再收到RT5和RT6的路由更新

RT7的路由表

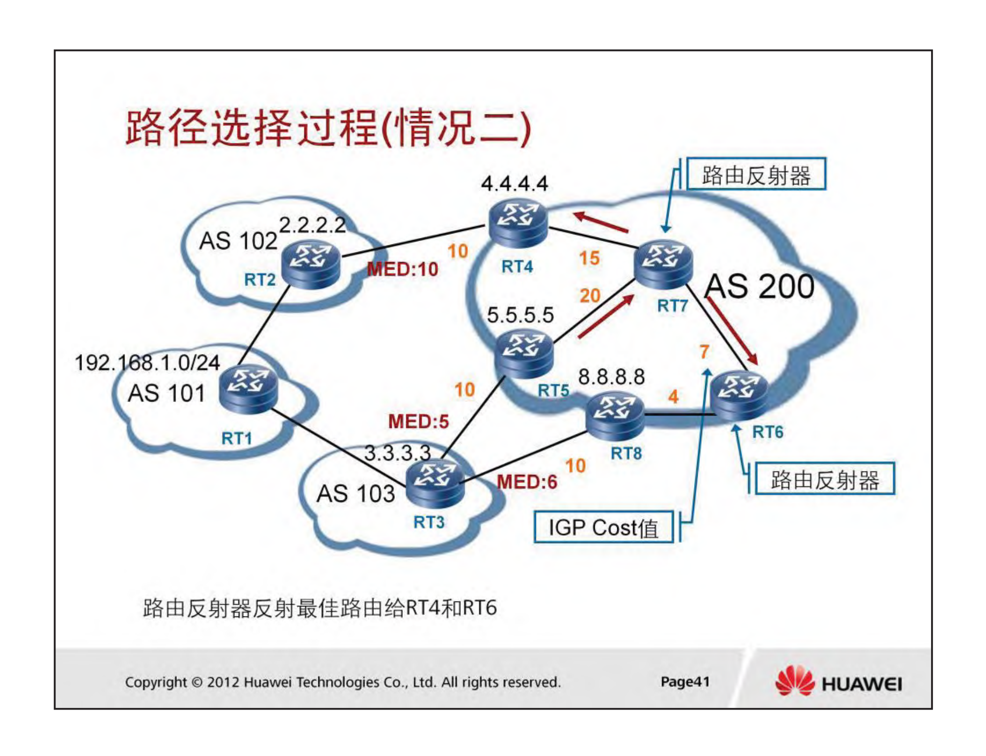
路径	BGP下一跳	AS-PATH	MED	IGP度量
1	RT4	102 101	10	15
2	RT8	103 101	6	11
3	RT5	103 101	5	20

 Page40



同样，根据之前所说。首先比较路径1和路径2的路由，路径2被选举；接着路径2与路径3比较，最终路径3被BGP选举为最佳路由。

由于RT7是路由反射器，选举后的结果会被反射到RT4和RT6上。


路径选择过程(情况二)
4.4.4.4
路由反射器
AS 102
2.2.2.2
10
RT4
15
RT2
MED:10
RT7
AS 200
20
5.5.5.5
7
192.168.1.0/24
8.8.8.8
AS 101
10
RT5
4
RT1
MED:5
RT8
RT6
3.3.3.3
10
路由反射器
AS 103
MED:6
RT3
IGP Cost值
路由反射器反射最佳路由给RT4和RT6
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page41
HUAWEI

路径选择过程(情况二)

RT4的路由表

路径	BGP下一跳	AS-PATH	MED	IGP度量
1 >	RT2	102 101	10	10
2	RT5	103 101	5	35

经过选举后，RT4仍然使用原来的路由条目

Page42



RT4与RT6都收到来自RT7路由反射信息。由于EBGP的优先级要比IBGP的优先级要高，所以RT4还是使用原来的路由条目。

路径选择过程(情况二)

RT6的路由表

路径	BGP下一跳	AS-PATH	MED	IGP度量
1 >	RT5	103 101	5	27
2	RT8	103 101	6	4

RT6选择从RT7反射来的路由信息，并回应UPDATE信息给RT7，撤消路由

RT6通过RT5作为下一跳，去往192.168.1.0/24

 Page43



RT6收到RT7的反射路由以及来RT8的路由之后，比较优劣，在具有相同AS_PATH属性的情况下，比较MED的大小，数值越小越优。最终选择RT7的反射路由。

一旦路径2被选择后，RT6马上回应一个UPDATE消息，目的是撤消(withdraw)原来发往RT7的路由信息。

路径选择过程(情况二)

收到RT6的UPDATE消息，RT7撤消路由后的路由表

RT7的路由表

路径	BGP下一跳	AS-PATH	MED	IGP度量
1 >	RT5	103 101	5	20
2	RT4	102 101	10	15
3	RT8	103 101	6	11

 Page44 

被撤消后的RT7路由表如图所示。

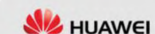
路径选择过程—故障处理

解决方案：把MED值重置为0

```
[RT4]route-policy med permit node 10
[RT4-route-policy]apply cost 0
[RT4]bgp 200
[RT4-bgp]peer 2.2.2.2 route-policy med import
```

```
[RT5]route-policy med permit node 10
[RT5-route-policy]apply cost 0
[RT5]bgp 200
[RT5-bgp]peer 3.3.3.3 route-policy med import
```

```
[RT6]route-policy med permit node 10
[RT6-route-policy]apply cost 0
[RT6]bgp 200
[RT6-bgp]peer 3.3.3.3 route-policy med import
```

 Page45 

从前面的例子我们可以看出，由于外部AS修改MED值并被传递到本地AS内，而BGP在默认的情况下并不对这类MED值进行修改，导致BGP选路过程中容易产生不可预测的结果，怎样解决这样的问题？就是把所有MED值重置为0，让MED值不再能影响BGP的选路，让内部IGP Cost影响BGP的选路，这就可以保证BGP从最近的出口转发数据到外部AS。

需要注意，这里举出的只是其中一种方法，由于BGP本来就是一个策略工具，在路径选择上可以有多种方法，这里就不一一举例了。

路径选择过程—故障处理

```
[RT7]display bgp routing-table

 Total Number of Routes: 3

 BGP Local router ID is 10.1.1.1
 Status codes: * - valid, > - best, d - damped,
               h - history,  i - internal, s - suppressed, S - Stale
               Origin : i - IGP, e - EGP, ? - incomplete
      Network            NextHop        MED        LocPrf    PrefVal Path/Ogn

 *>i  192.168.1.0/24     6.6.6.6                    100        0      103 101i
 * i                     4.4.4.4                    100        0      102 101i
 * i                     5.5.5.5                    100        0      103 101i
```


Page46



路径选择过程—总结

由于BGP的路径选择参数很多，很容易因为某参数的不正确配置导致BGP路由表产生不正常的现象。主要影响BGP路径选择的参数有：

- EBGP对等体之间
  - AS_PATH
  - MED
  - ORIGINATOR_ID / ROUTER_ID
- IBGP对等体之间
  - IGP Cost值
  - MED

 Page47



由于BGP的路径选择参数很多，很容易因为某参数的不正确配置导致BGP路由表产生不正常的现象。主要影响BGP路径选择的参数有：

- EBGP对等体之间
  - AS_PATH
  - MED
  - ORIGINATOR_ID / ROUTER_ID
- IBGP对等体之间
  - IGP Cost值
  - MED

当然除了以上参数以外，BGP的其它参数值都能影响BGP的选路，例如：community属性，Local_Pref等。





1.BGP邻居关系的建立需要注意什么？

答：BGP邻居关系的建立需要注意TCP端口179是否被禁用，邻居之间是否存在IP连通性的问题。EBGP/IBGP的建立需要注意的事项，如：EBGP多跳问题，IBGP更新源地址问题。OPEN消息的参数如：自治系统号是否匹配，BGP ROUTER ID的配置等。BGP人为配置错误也会导致BGP邻居关系无法建立。

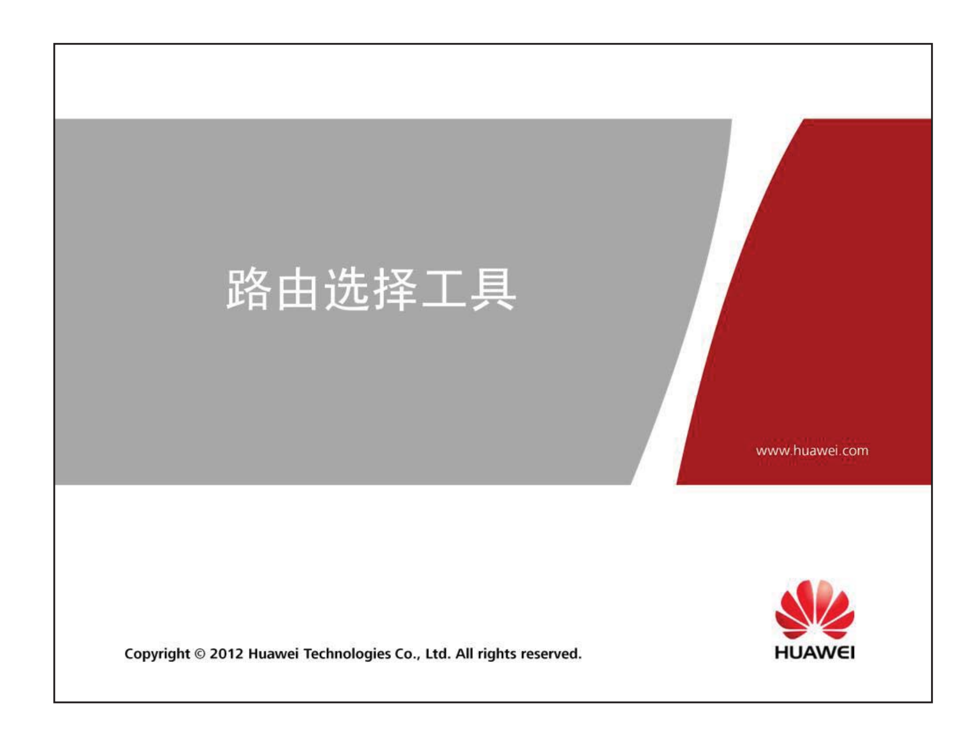
2.BGP路由学习需要注意什么？

答：当BGP邻居关系建立以后，邻居之间就开始交互UPDATE消息学习路由信息。这里最需要注意的就是BGP的通告原则，通过“network”命令通告的前缀必须存在IP路由表中，而且需要严格匹配其掩码长度；通过“aggregate”命令通告的必须存在于BGP路由表中，否则无效。另外就是IBGP下一跳不可达问题，建议通过命令“peer next-hop-local”把下一跳修改为本地。


谢谢
www.huawei.com

Module 4

路由选择和控制


路由选择工具
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

要执行定义的策略，必须先确定执行的对象，此时我们可以使用路由选择工具。

Page1





培训目标

学完本课程后，您应该能：

- 理解路由选择的各种工具及其作用

Page2



路由选择工具

访问控制列表（ACL）

- 用于匹配路由信息或者数据包的地址，过滤不符合条件的路由信息或数据包

前缀列表（ip-prefix）

- 匹配对象为路由信息的目的地址或直接作用于路由器对象（gateway）

自治系统路径信息访问列表（ as-path-filter）

- 仅用于BGP协议，匹配BGP路由信息的自治系统路径域

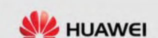
团体属性列表（ community-filter）

- 仅用于BGP协议，匹配BGP路由信息的自治系统团体域

路由策略（route-policy）

- 设定匹配规则，由if-match和apply子句组成

 Page3



当我们需要执行路由策略或者策略路由的时候，一般先要把特定的路由信息或者数据包过滤出来。可以根据过滤不同的对象采用不同的过滤工具。访问控制列表和前缀列表一般都可以用来匹配IP地址，但是前缀列表不能用来过滤数据包，只能用来过滤路由信息。所以我们要首先清楚匹配的对象是什么，是路由还是数据，然后才能选择适当的工具。

As-PATH-FILTER是用来匹配BGP路由信息中的AS-PATH属性的，所以它只能用于过滤BGP路由。

Community-filter是用来匹配BGP路由信息中的团体属性的，所以，同as-path-filter一样只能用于过滤BGP路由。

Route-POLICY是一个强大的过滤工具，但是它还是策略工具。作为过滤工具，它可以用if-match语句来匹配路由和数据包，而且if-match语句还可以调用其它过滤工具。作为策略工具，它可以使用apply语句来修改路由属性或者数据包的转发行为。

访问控制列表

访问控制列表是由**permit | deny**语句组成的一系列有顺序的规则，这些规则根据源地址、目的地址、端口号等来描述。

按照访问控制列表的用途，可以分为三类：

- 基本的访问控制列表（basic acl）
- 高级的访问控制列表（advanced acl）
- 基于接口的访问控制列表（interface-based acl）

访问控制列表的使用用途是依靠数字的范围来指定的

- 2000～2999：基本的访问控制列表
- 3000～3999：高级的访问控制列表
- 1000～1999：基于接口的访问控制列表

 Page4



访问控制列表既可以用来匹配数据包也可以用来匹配路由信息。

基本访问控制列表可以用来匹配源IP地址。

高级访问控制列表可以用来匹配源IP地址、目的IP地址、源端口号、目的端口号、协议号等。

基于接口的访问控制列表可以用来匹配接口。

访问控制列表的匹配顺序

有两种匹配顺序：

- 配置顺序
  - 配置顺序，是指按照用户配置ACL规则的先后进行匹配
- 自动排序
  - 自动排序使用“深度优先”的原则
  - “深度优先”规则是把指定范围最小的语句排在最前面

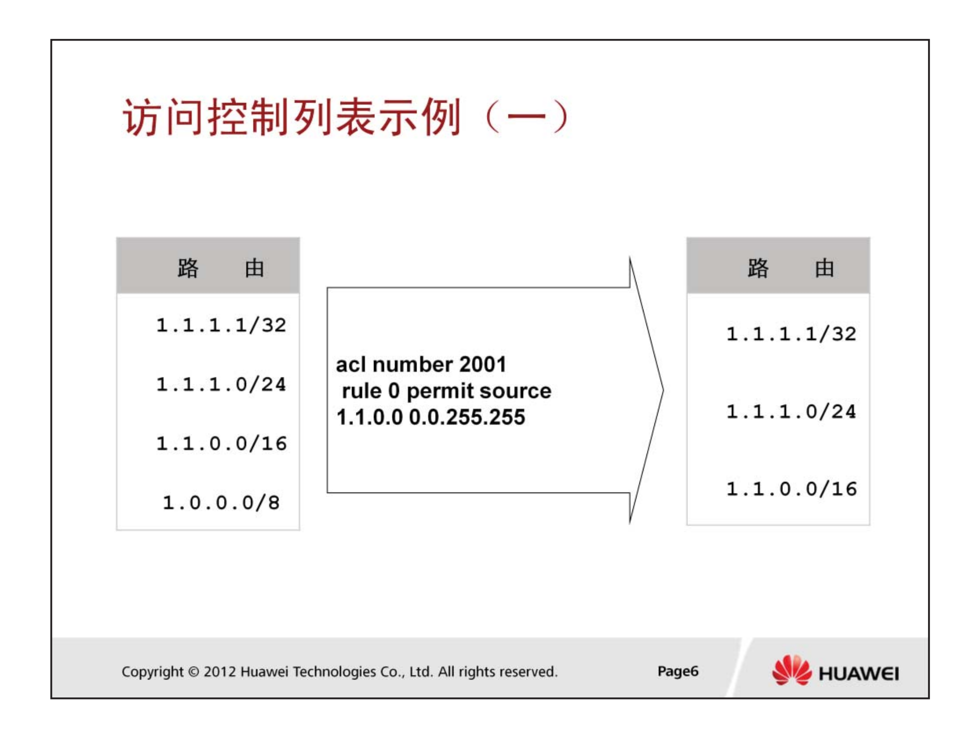
 Page5 

有两种匹配顺序：配置顺序和自动排序。

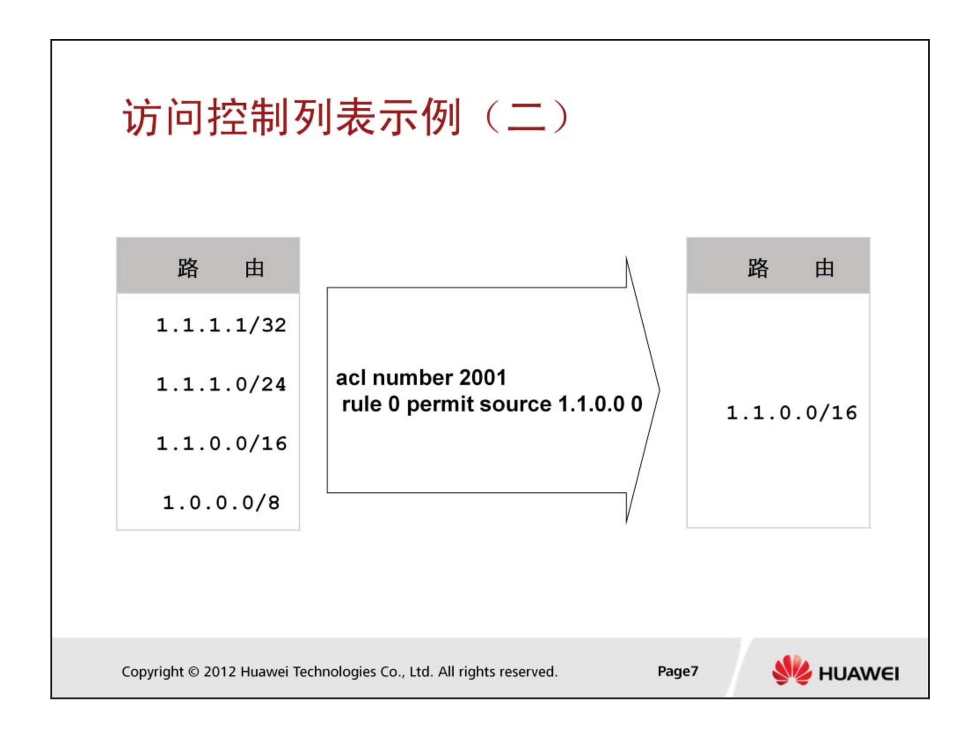
配置顺序，是指按照用户配置ACL规则的先后进行匹配。

自动排序使用“深度优先”的原则。

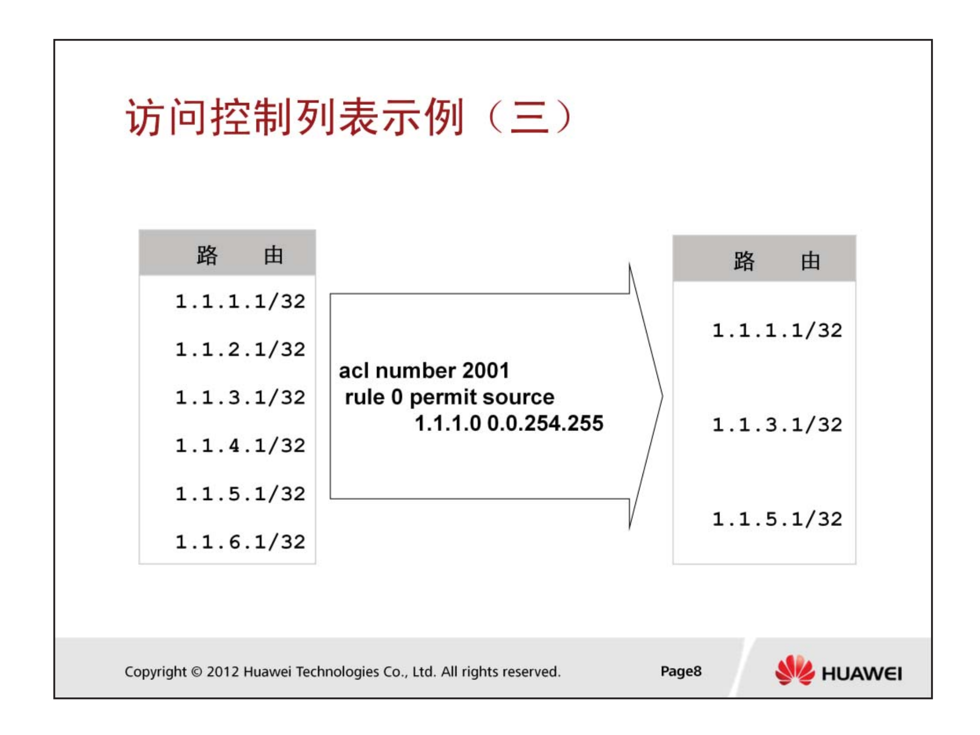
“深度优先”规则是把指定数据包范围最小的语句排在最前面。这一点可以通过比较地址的通配符来实现，通配符越小，则指定的主机的范围就越小。比如129.102.1.1 0.0.0.0指定了一台主机：129.102.1.1，而129.102.1.1 0.0.0.255则指定了一个网段：129.102.1.1～129.102.1.255，显然前者在访问控制规则中排在前面。具体标准为：对于基本访问控制列表，直接比较源地址通配符，通配符相同的则按配置顺序；对于基于接口的访问控制列表，配置了“any”的规则排在后面，其它按配置顺序；对于高级访问控制列表，首先比较源地址通配符，相同的再比较目的地址通配符，仍相同的则比较端口号的范围，范围小的排在前面，如果端口号范围也相同则按配置顺序。



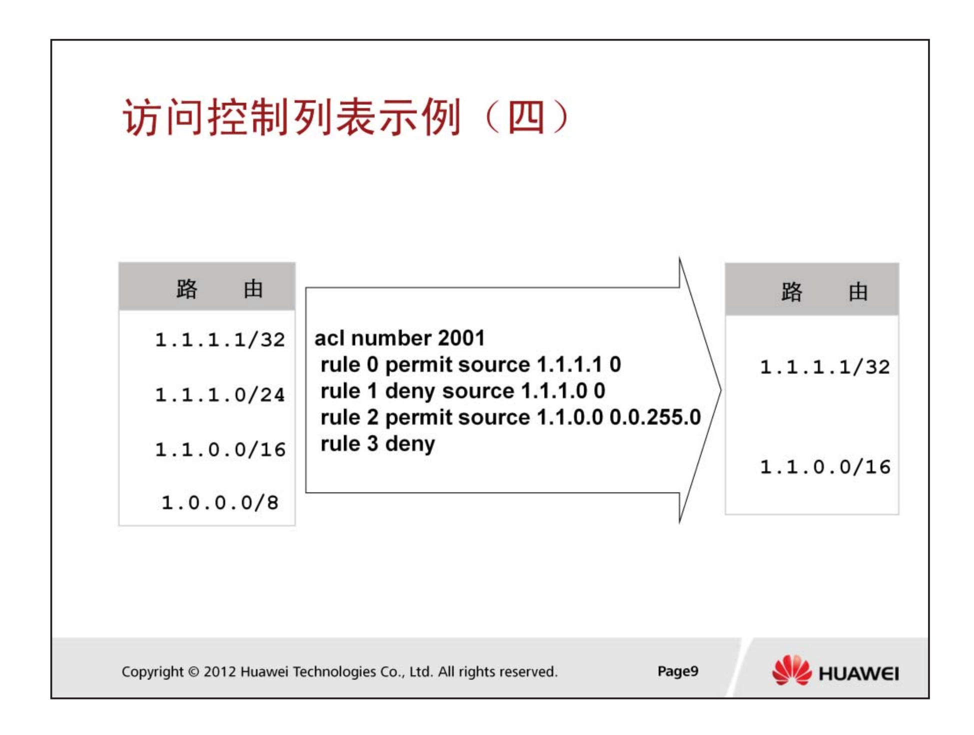
这个ACL的匹配条件是“1.1.0.0 0.0.255.255”，意思是只要路由的前两个字节是“1.1”就能满足匹配条件，后两个字节不影响匹配的结果。因此，“1.1.1.1/32”，“1.1.1.0/24”和“1.1.0.0/16”都满足匹配条件。



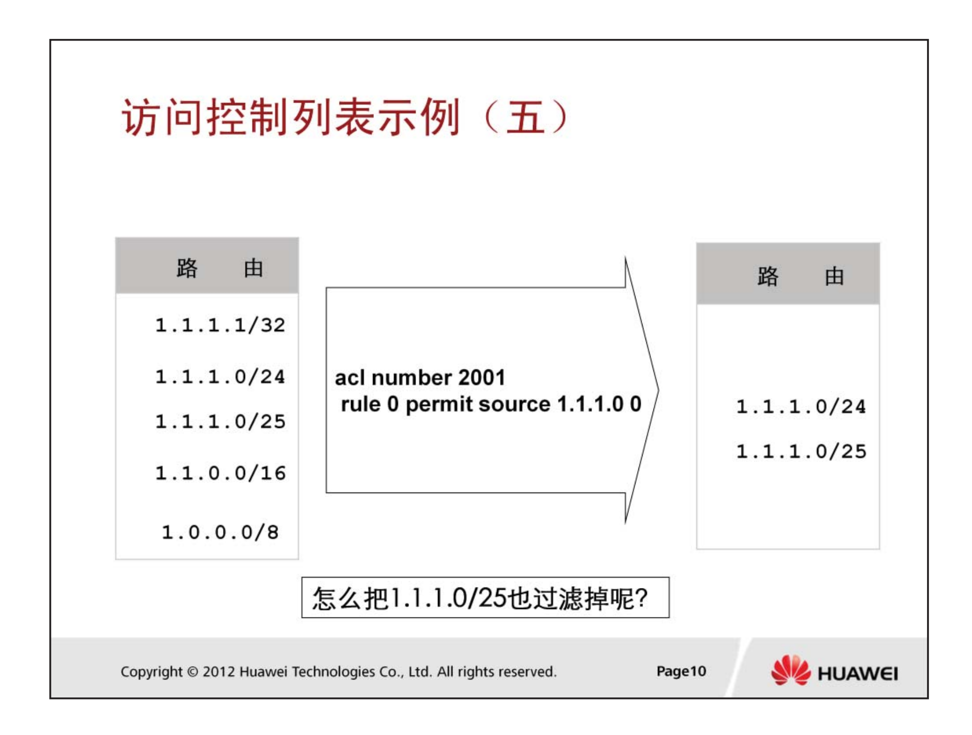
这个ACL的匹配条件是“1.1.0.0 0”，其中反掩码是“0”，这意味着路由条目的所有32个bits必须是“1.1.0.0”，因此只有“1.1.0.0/16”这个路由条目满足匹配条件。



这个示例中的匹配条件是“1.1.1.0 0.0.254.255”，其中反掩码的二进制表示形式为“00000000.00000000.11111110.11111111”，请记住：反掩码中，0表示严格匹配，1表示不关心。所以，这个匹配条件表明，前两个字节必须严格匹配，第三个字节的前7位不关心，第8位必须严格匹配，第四个字节不关心。将“1.1.1.0”跟反掩码相比较，我们可以得出结论：这个匹配条件所匹配的路由的前两个字节必须是“1.1”，第3个字节的最后一个bit必须是1（表明这个字节是个奇数）。所以，示例中满足这个条件的路由有“1.1.1.1/32，1.1.3.1/32，1.1.5.1/32”，其它的路由条目都不满足第三个字节是奇数的条件。



在一个ACL中可以同时定义多个过滤条件，在这个示例中，我们给ACL 2001定义了4个过滤条件。“1.1.1.1/32”匹配“rule 0”；“1.1.1.0/24”匹配”rule 1”，因此被过滤掉了，1.1.0.0/16”满足“rule 2”；“1.0.0.0/8”不满足前三个匹配条件，因此被最后的“rule 3”给过滤掉了。



使用ACL可以很好匹配路由的前缀部分，但是对于前缀相同，掩码不同的路由怎么区分呢？这时候，可以使用前缀列表。

前缀列表

前缀列表用来过滤IP前缀，能同时匹配前缀号和前缀长度

前缀列表的性能比访问控制列表高

前缀列表不能用于数据包的过滤

例：**ip ip-prefix test index 10 permit 10.0.0.0 16 greater-equal 24 less-equal 28**

- 前缀号必须为10.0
- 24<=前缀长度<=28
- 满足条件的如10.0.1.0/24, 10.0.2.0/25, 10.0.2.192/26

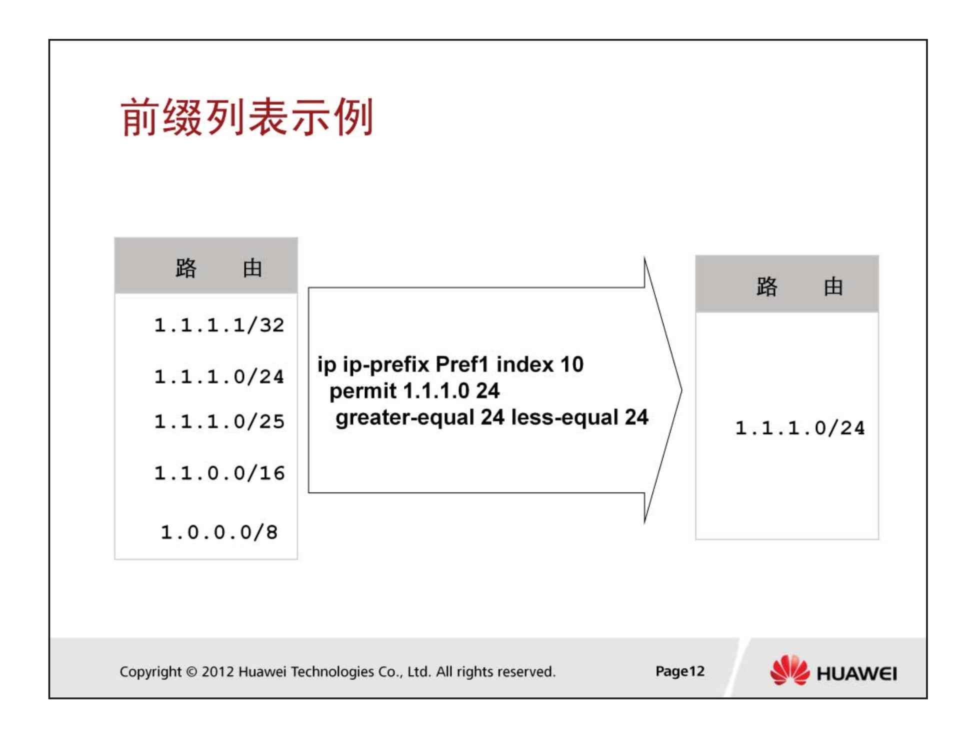
Page11



前缀列表用来过滤IP前缀，能同时匹配前缀号和前缀长度；前缀列表的性能比访问控制列表高；前缀列表不能用于数据包的过滤。

例：ip ip-prefix test index 10 permit 10.0.0.0 16 greater-equal 24 less-equal 28

- 前缀号必须为10.0
- 24<=前缀长度<=28
- 满足条件的如10.0.1.0/24, 10.0.2.0/25, 10.0.2.192/26



在这个前缀列表中， “index 10”定义了两个匹配条件：一个是“1.1.1.0 24 ”；另一个是“greater-equal 24 less-equal 24”。

“1.1.1.0 24 ”表示路由的前缀部分的前三个字节必须是“1.1.1”；“greater-equal 24 less-equal 24”表示路由的掩码长度必须是24位。所以，只有“1.1.1.0/24”这条路由满足匹配条件。

另外需要注意的是，前缀列表可以同时定义多个“index”。

AS-Path-Filter

该列表用来过滤BGP的AS-PATH属性

AS-PATH属性使用正则表达式来定义

- 举例
  - 匹配所有AS-PATH属性 ip as-path-filter 10 permit .*
  - 匹配从AS100发起的路由 ip as-path-filter 10 permit _100$
  - 匹配从AS200接收的路由 ip as-path-filter 10 permit ^200_

 Page13



AS-PATH属性被用来记录路由在传递过程中经过的AS号。如果一条路由起源于AS100，然后一次经过AS300, AS200, AS500，最后到达AS600。那么在AS600里，路由的AS-PATH属性表示为‘500 200 300 100’。AS-PATH属性实际上是一个字符串，因此可以用正则表达式来表示。

AS-Path-Filter和正则表达式

字符	符号	特殊意义
句号	.	匹配任意单字符
星号	*	匹配模式中0或更多序列
加号	+	匹配模式中1或更多序列
问号	?	匹配模式0或一次出现
加字符	^	匹配输入字符串的开始
美元符	$	匹配输入字符串的结束
下划线	_	匹配逗号，括号，字符串的开始和结束，空格
方括号	[范围]	表示一个单字符模式的范围
连字符	-	把一个范围的结束点分开

 Page14 HUAWEI

正则表达式是非常灵活的，同样一种含义可以有多种表示方法。

Community-filter

团体列表使用团体属性表示和过滤BGP路由

团体列表有基本和高级两种

- 基本团体列表用来匹配实际的团体属性值和常量
  - ip community-filter 1 permit 100:1 100:2
  - ip community-filter 1 permit 100:1
  - ip community-filter 1 permit no-export
- 高级团体列表可以使用正则表达式
  - ip community-filter 100 permit ^10

Page15



团体属性分为Well-known团体属性和私有的团体属性。

Well-known团体属性包括：internet, no-advertise, no-export, no-export-subconfed。

私有团体属性由管理员定义，主要用来给前缀打上管理标记，以便制定相应的策略，格式一般为AS:NUMBER。

高级团体列表可以使用正则表达式来匹配团体属性。

Route-policy

一个routing policy下可以有多个节点，不同的节点号用seq-number标识，不同seq-number各个部分之间的关系是“或”的关系

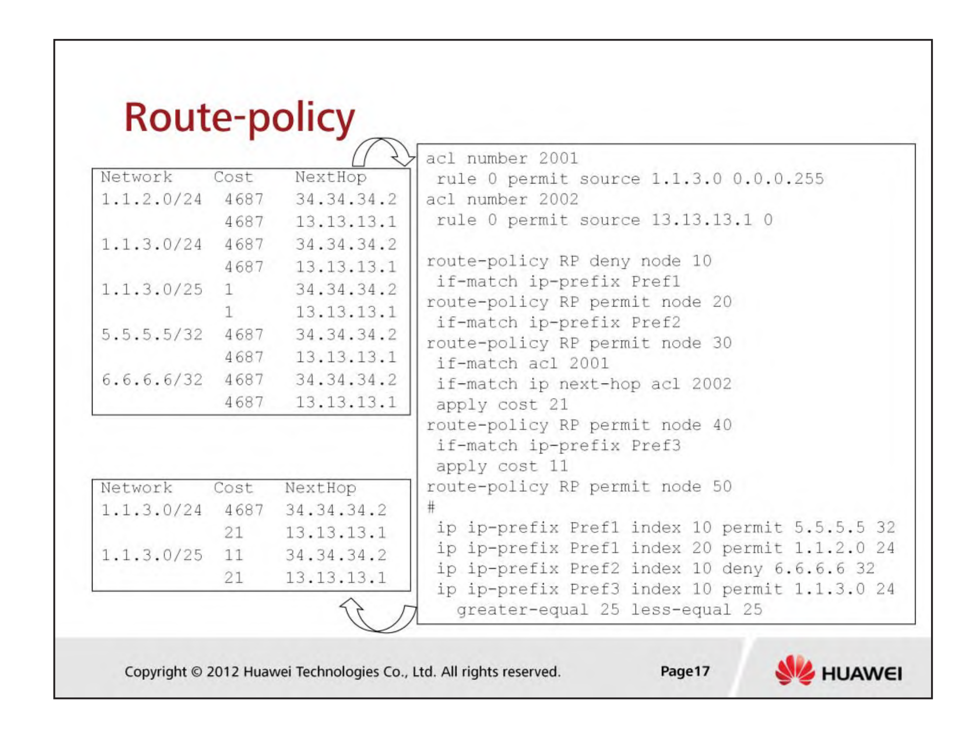
- 每个节点下可以有多个if-match和apply子句，if-match子句之间是“与”的关系
- 允许模式：当路由项满足该节点的所有if-match子句时，将被允许通过该节点的过滤并执行该节点的apply子句，如路由项不满足该节点的if-match子句，将试图匹配路由策略的下一个节点
- 拒绝模式：当路由项满足该节点的所有if-match子句时，将被拒绝通过该节点的过滤，并且不会继续下一个节点的测试
- If-match子句可以引用其它的过滤工具

　Page16



Route-policy是强大的过滤工具和策略工具。它由一系列的if-match子句和Apply子句构成。If-match子句可以用来匹配acl, ip-prefix, as-path-filter, community-filter, interface, ip, extcommunity-filter, cost, mpls-label, route type, tag等。Apply子句可以用来修改路由的属性。

Route-policy可以用于路由策略。



路由表中的路由按照顺序依次匹配路由策略的节点。

首先，定义了前缀列表Pref1，用来匹配5.5.5.5/32和1.1.2.0/24，匹配此前缀列表的条目将被路由策略的node 10过滤掉（deny），所以在过滤后的路由表中看不到5.5.5.5/32和1.1.2.0/24；

前缀列表Pref2用来过滤6.6.6.6/32（deny），所以尽管路由策略的node 20的策略是permit，6.6.6.6/32仍然被过滤掉了；

路由策略的node 30定义了两个if-match语句，件还同时满足匹配acl 2001和下一跳满足acl 2002的条件，acl 2001匹配的路由条目包括1.1.3.0/24和1.1.3.0/25共4条路由，但是满足node 30定义的条需要满足下一跳为13.13.13.1，因此只有两条路由可以满足条件，这两条路由的cost被apply语句修改为21；

剩余的两条路由是1.1.3.0/24和1.1.3.0/25（下一跳是34.34.34.2），这两条路由继续试图尝试匹配路由策略的node 40，于是1.1.3.0/25路由可以匹配前缀列表Pref3，cost被修改为11；

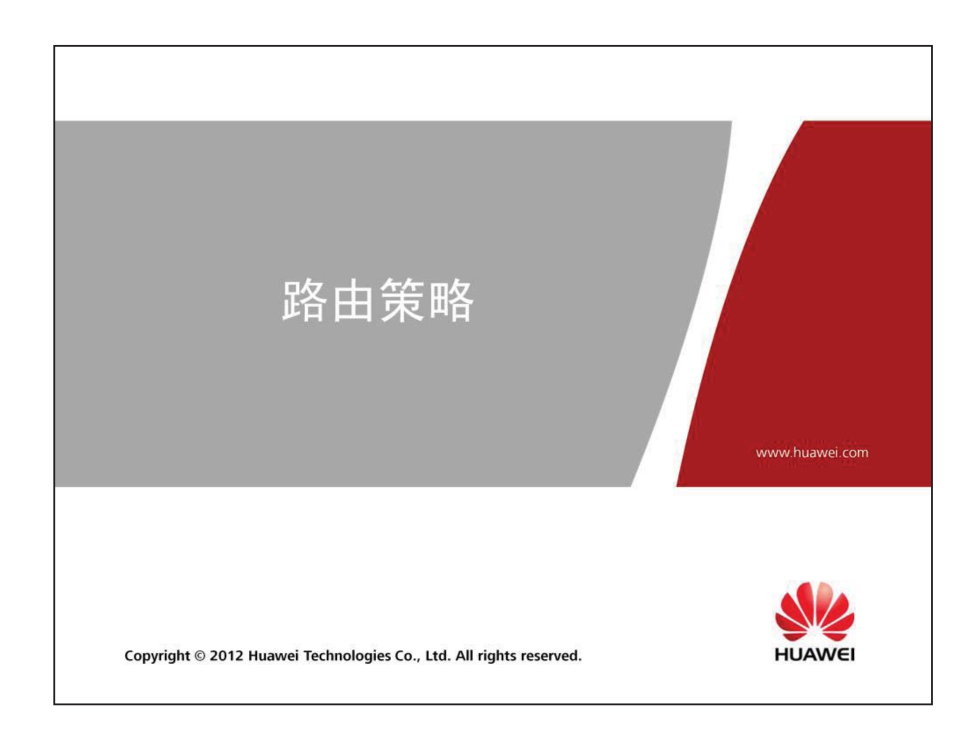
最后剩下的路由1.1.3.0/24（下一跳是34.34.34.2 ）被node 50原封不动的保留。



Q：本课程介绍了哪些路由选择工具?

A：访问控制列表；前缀列表；AS-PATH-FILTER；COMMUNITY-FILTER；ROUTE-POLICY。


谢谢
www.huawei.com


路由策略
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI





前 言

我们可以使用路由策略来控制路由的选择和引入过程。

Page1





培训目标

学完本课程后，您应该能：

- 使用路由策略控制路由的接收过程
- 使用路由策略控制路由的引入过程
- 使用路由协议优先级来选择路由

Page2





路由引入

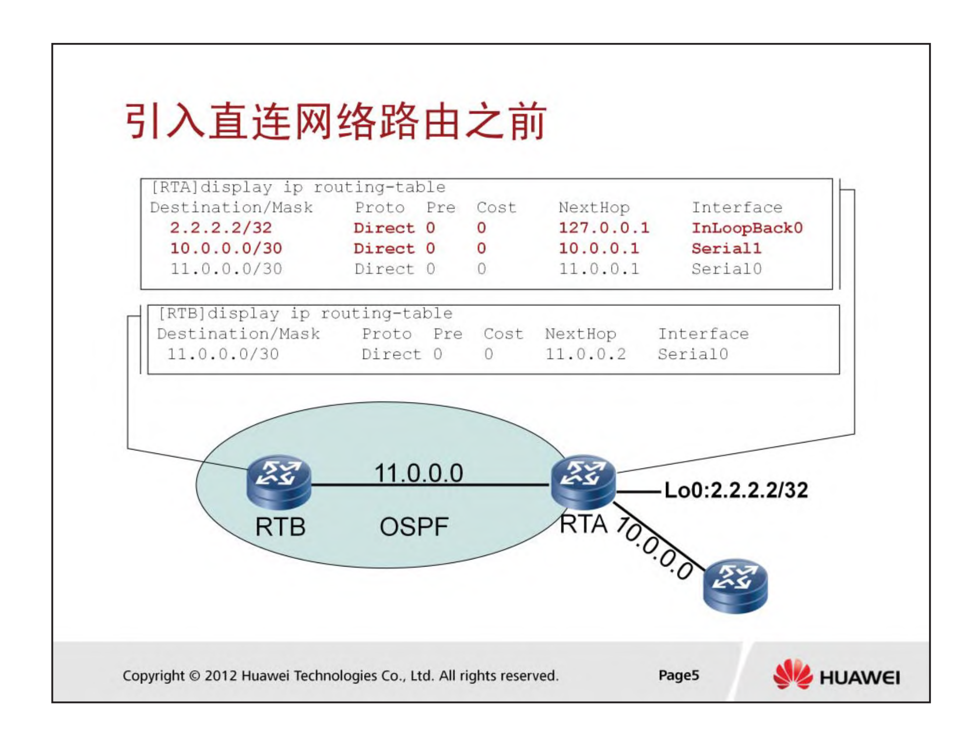
路由过滤

使用路由协议优先级

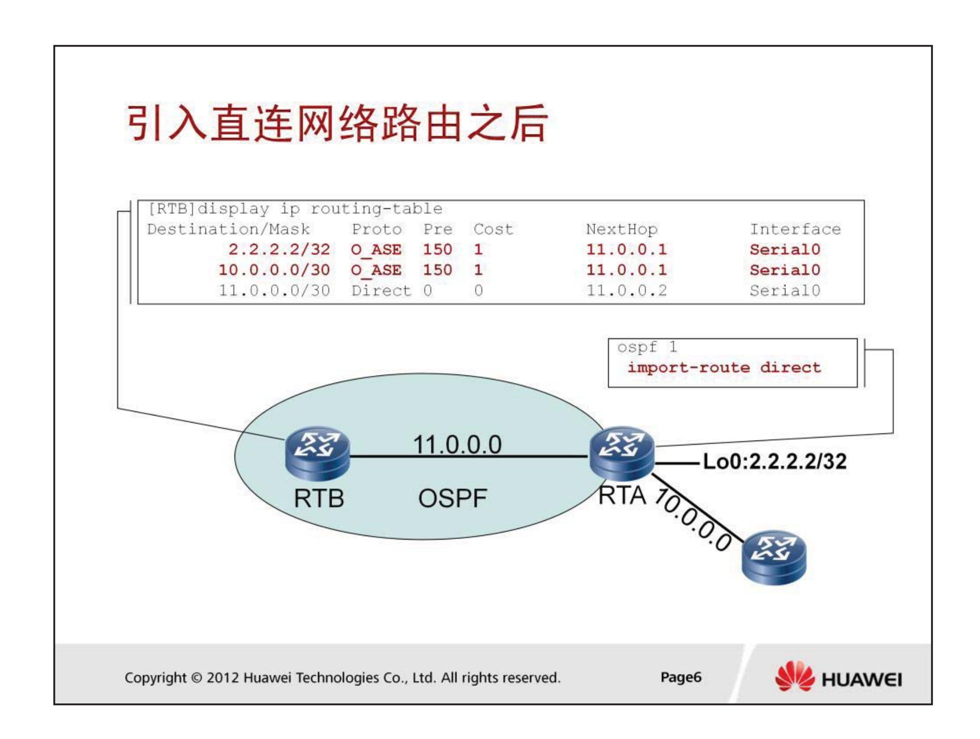
控制缺省路由下发

Page4

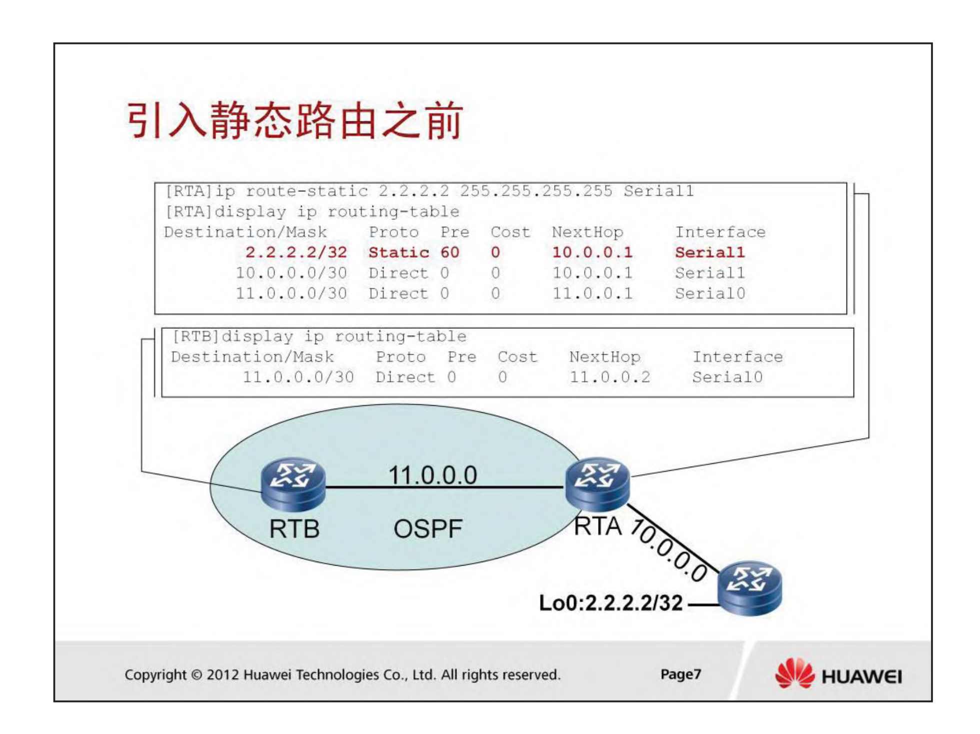




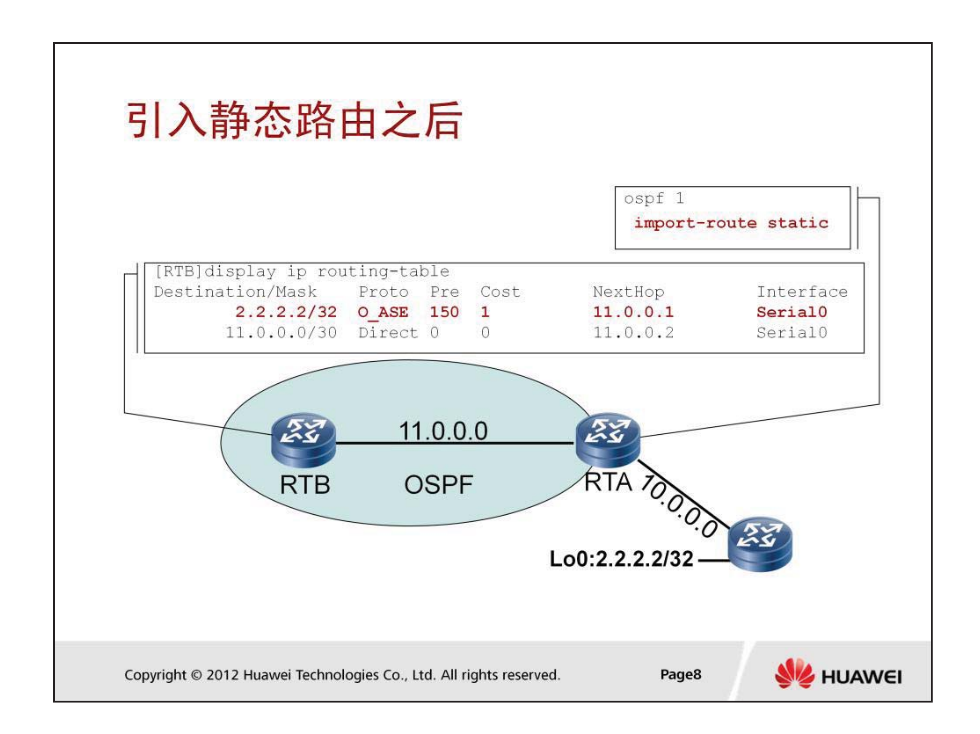
在RTA上，只有11.0.0.0网段运行OSPF，在它的路由表中，2.2.2.2和10.0.0.0网段都是直连网段，在没有配置路由引入的时候， 2.2.2.2和10.0.0.0网段是不会通告给RTB的，在RTB的路由表中没有任何关于2.2.2.2和10.0.0.0网段的信息。



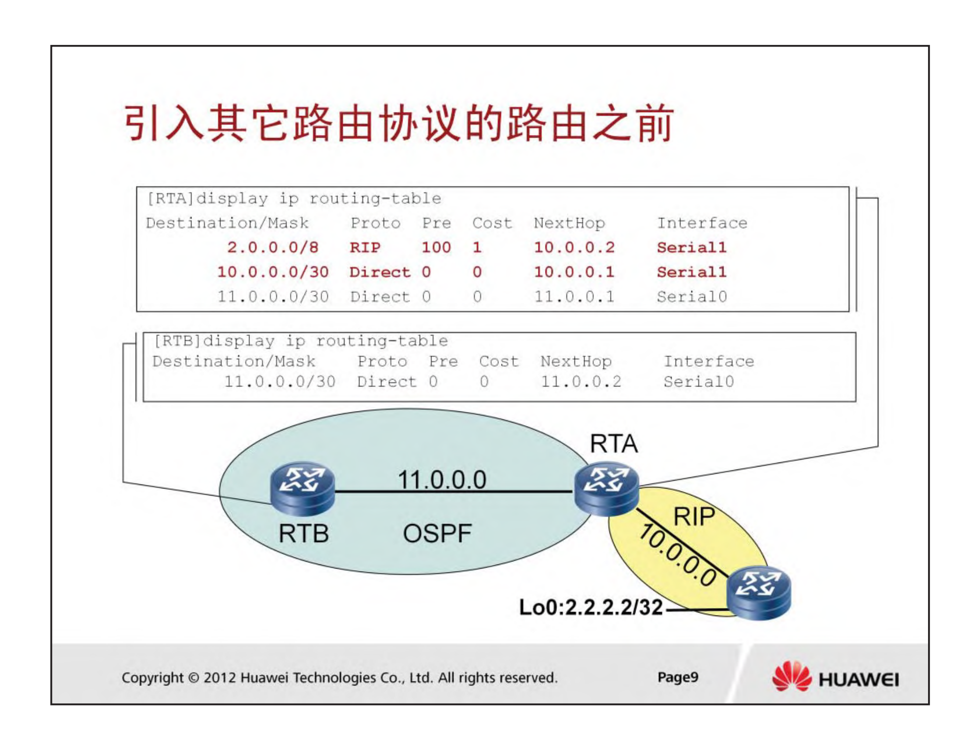
在RTA上把直连路由引入到OSPF中，这时，RTA就会把2.2.2.2和10.0.0.0网段的路由信息在OSPF进程中通告给RTB，因此，在RTB上，我们可以看到2.2.2.2和10.0.0.0网段的路由，而且它们是通过OSPF学习到的（Proto字段标记为O_ASE，即OSPF外部路由）。



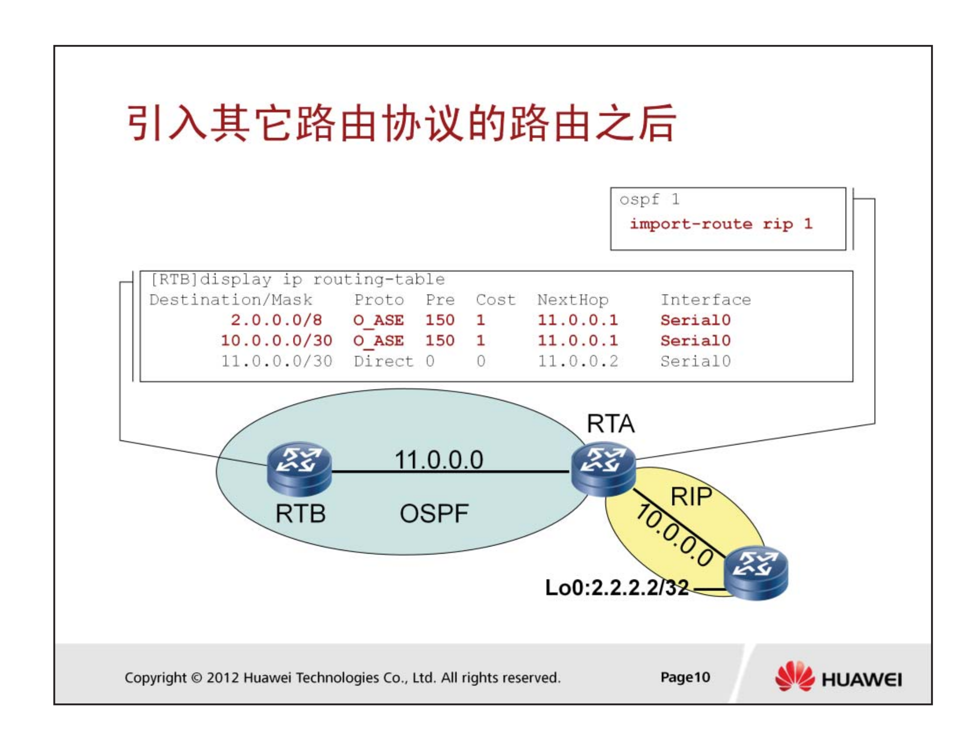
在RTA上，我们配置了一条到2.2.2.2网段的静态路由。如果在RTA上没有配置路由引入，这条静态路由是不会自动通告到OSPF进程中的。



在RTA配置引入静态路由，这时，2.2.2.2网段就会通过OSPF进程被通告给RTB，而10.0.0.0网段既没有运行OSPF，也不是静态路由，所以不会通告给RTB。



在这个网络中，同时运行OSPF和RIP，运行OSPF的网段包括11.0.0.0，运行RIP的网段包括10.0.0.0和2.0.0.0。在不配置路由引入的情况下，RTB没有任何关于2.0.0.0和10.0.0.0网段的路由信息。



在RTA上把RIP引入到OSPF，RTB就可以学到2.0.0.0和10.0.0.0网段的路由信息。注意，在RTA的路由表中，尽管10.0.0.0网段是直连网段，但是，因为这个网段运行了RIP，所以也会被引入到OSPF中。

为什么需要配置路由引入

部署不同路由协议的机构合并

不同的网络使用不同的协议，并且这些网络需要共享路由信息

- 简单的网络可以使用RIP
- 网络类型复杂的可以选用OSPF
- 大型骨干网络一般选用ISIS

网络协议的限制

- 比如，使用拨号链路连接两个ISIS网络，而在拨号链路上是不适合运行ISIS协议的
- 需要配置静态路由，然后把静态路由引入到ISIS

 Page11



在一般情况下，部署一种路由协议就够了。但是，在以下这些情况，我们可能要部署路由引入。

1、部署不同路由协议的机构合并

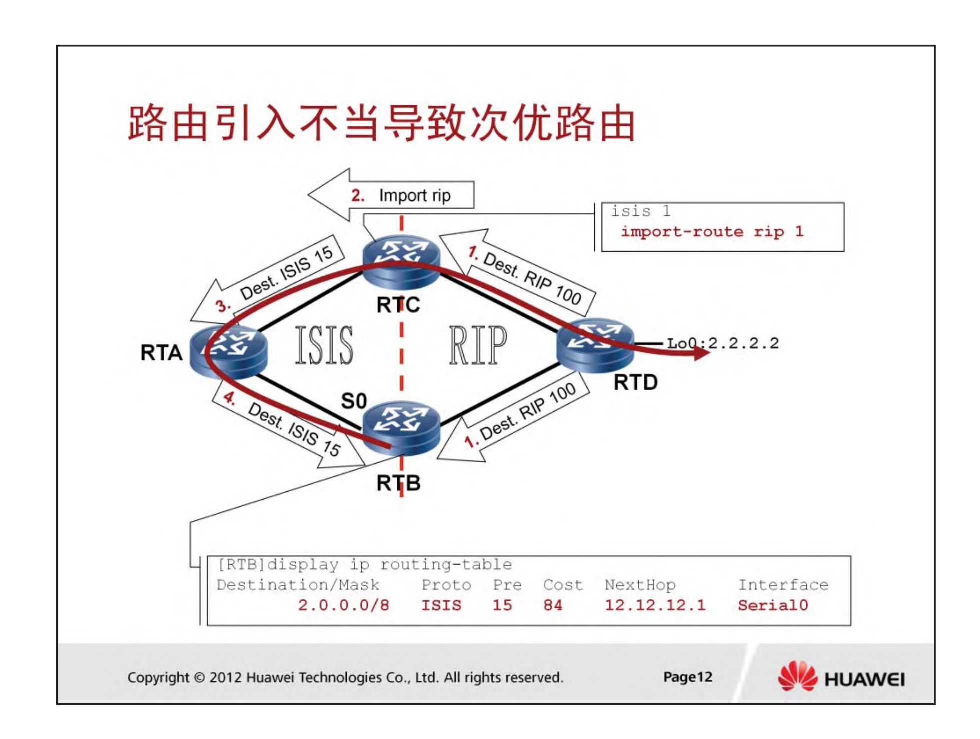
比如，A公司部署OSPF，B公司部署ISIS，现在A公司和B公司合并成一个C公司，这时候，原来A公司的网络要和原来B公司的网络互相访问，需要配置路由引入。

2、不同的网络使用不同的协议，并且这些网络需要共享路由信息

一个很大的网络可能由很多小网络组成，这些小网络的复杂度是不一样的，有些网络很小，为了管理简单，所以部署了RIP，有些网络的链路类型很复杂，所以部署OSPF（OSPF比ISIS支持的网络类型多），而其它网络部署ISIS。为了实现这些小网络的互访，可能要配置路由引入。

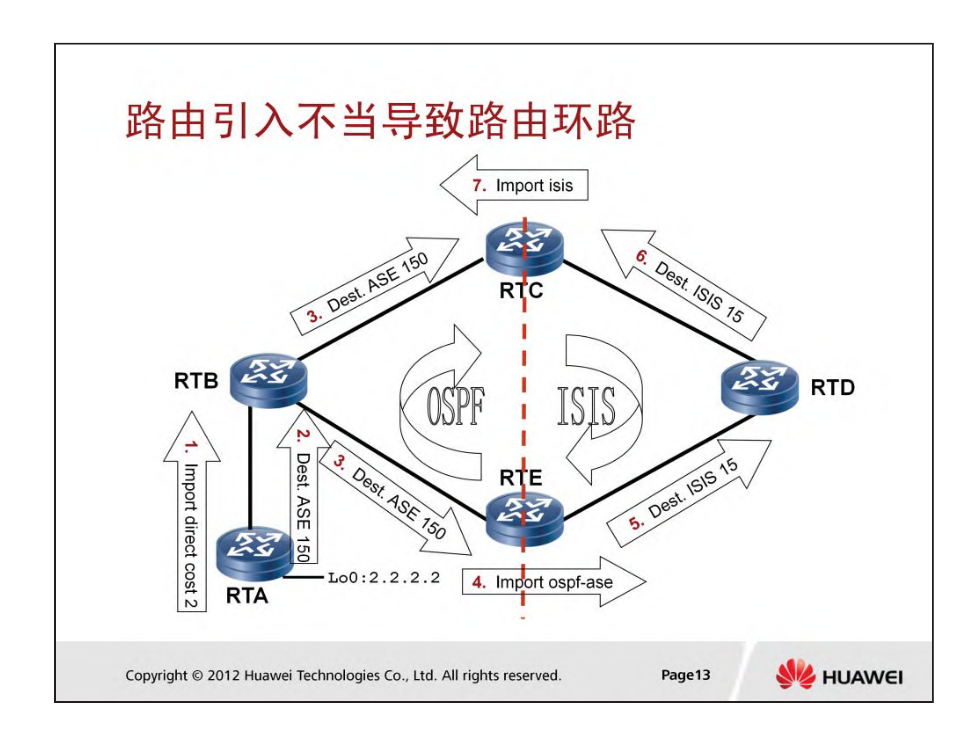
3、网络协议的限制

拨号网络是按时间计费的，所以拨号网络一般只作为备份链路使用，在主链路正常的情况下，拨号链路是不工作的。在拨号网络上，不适合运行ISIS协议（OSPF协议有专门针对拨号链路的设计），因为ISIS协议需要定时发送报文，如果在拨号链路上运行ISIS协议，这些报文会导致拨号链路在主链路正常的情况下也处于UP状态。一般地，我们在拨号链路上配置静态路由，然后把静态路由引入到ISIS中。



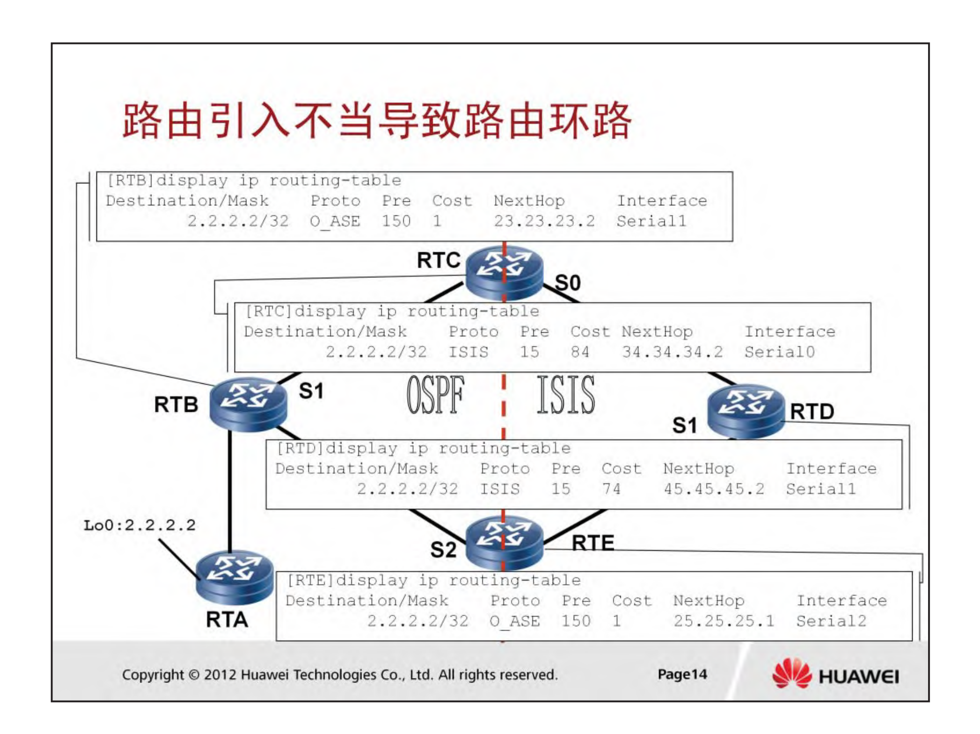
在这个网络中，同时运行ISIS和RIP协议。其中RTC和RTB是ASBR。在RTC上，我们把RIP引入到ISIS中。2.2.2.2这个网段将通过ISIS通告给RTA，然后RTA又通告给RTB。RTB同时从RIP和ISIS学习到关于2.2.2.2的路由，于是它比较ISIS和RIP的优先级，因为ISIS的优先级为15，RIP的优先级为100，所以RTB最终选择ISIS通告的路由。于是，RTB如果要把数据包发往2.2.2.2的话，将选用RTB-RTA-RTC-RTD的次优路径。

因此，在配置路由引入的时候，要避免次优路由的产生。如何避免，后面的课程会有讨论。

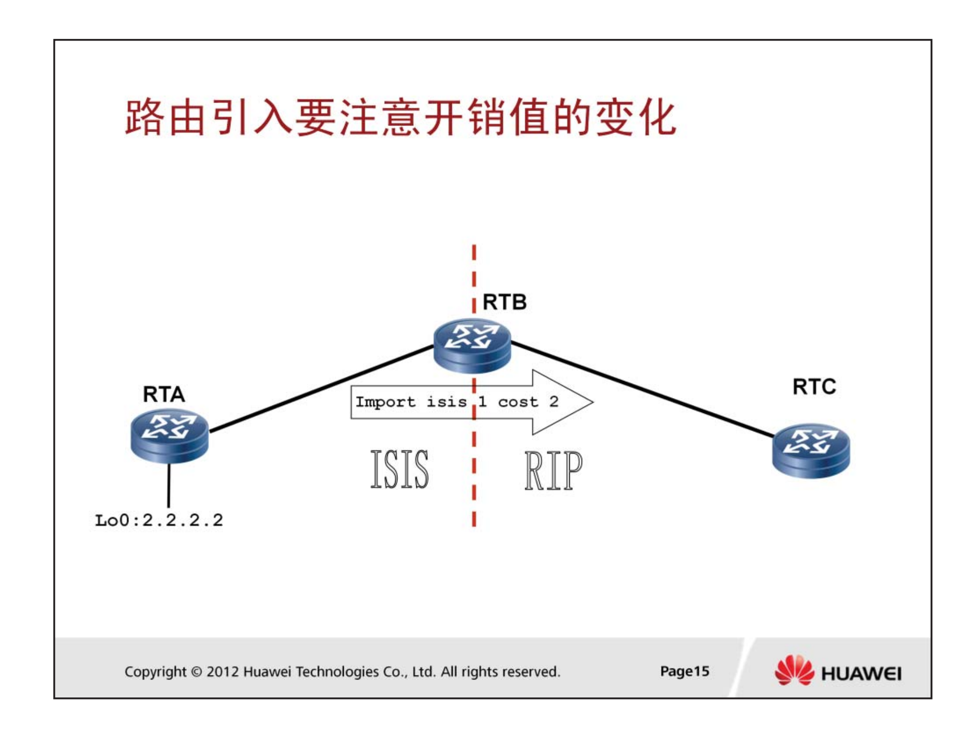


RTA通过引入直连路由把2.2.2.2网段引入到OSPF中，OSPF将采用ASE路由（优先级为150）的方式通告给RTB, RTC, RTE。在RTE上，配置了引入OSPF-ASE，把2.2.2.2引入到ISIS中。在RTC上，配置了引入ISIS，把ISIS路由引入到OSPF中，于是， 2.2.2.2网段又从ISIS通告回OSPF，这叫做路由回馈。这样的话，RTB同时从RTA和RTC学习到关于2.2.2.2网段的路由，因为优先级都一样（都是OSPF ASE路由），所以比较metric值，如果RTB很不幸地选择了RTC通告的路由，环路就产生了。比如：RTD发一个数据包到2.2.2.2 ，数据包将发往RTE，然后到RTB，因为RTB选择了RTC的路由，所以RTB把数据包发往RTC，RTC再发到RTD，于是数据包回到了起点。

在复杂的环境中要小心避免这种情况的出现，如何避免，后面课程会讲述。



从路由器的路由表可以看出，环路已经产生了。



不同的路由协议计算路由开销的依据是不同的，开销值的大小是不同的，而且开销的范围也是不同的。ISIS和OSPF的开销值可以基于带宽，而且值的范围很大，RIP的开销基于跳数，范围很小，所以，当我们配置ISIS和RIP的引入或者OSPF和RIP的引入的时候一定要小心（在VRP平台上，当我们引入OSPF或者ISIS路由到RIP的时候，如果不指定COST，开销值将默认设为1，尽管如此，我们还是应该手工配置开销值以反映网络的真实拓扑）。



目 录

Page16



路由过滤的作用

避免路由引入导致的次优路由

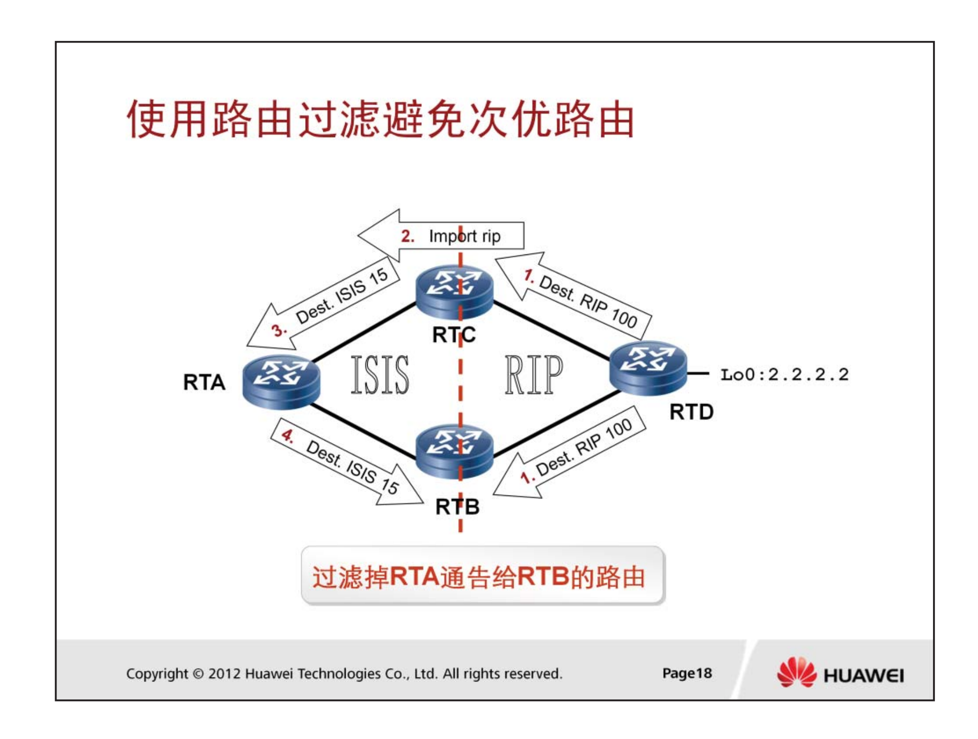
避免路由回馈导致的路由环路

进行精确的路由引入和路由通告控制

 Page17

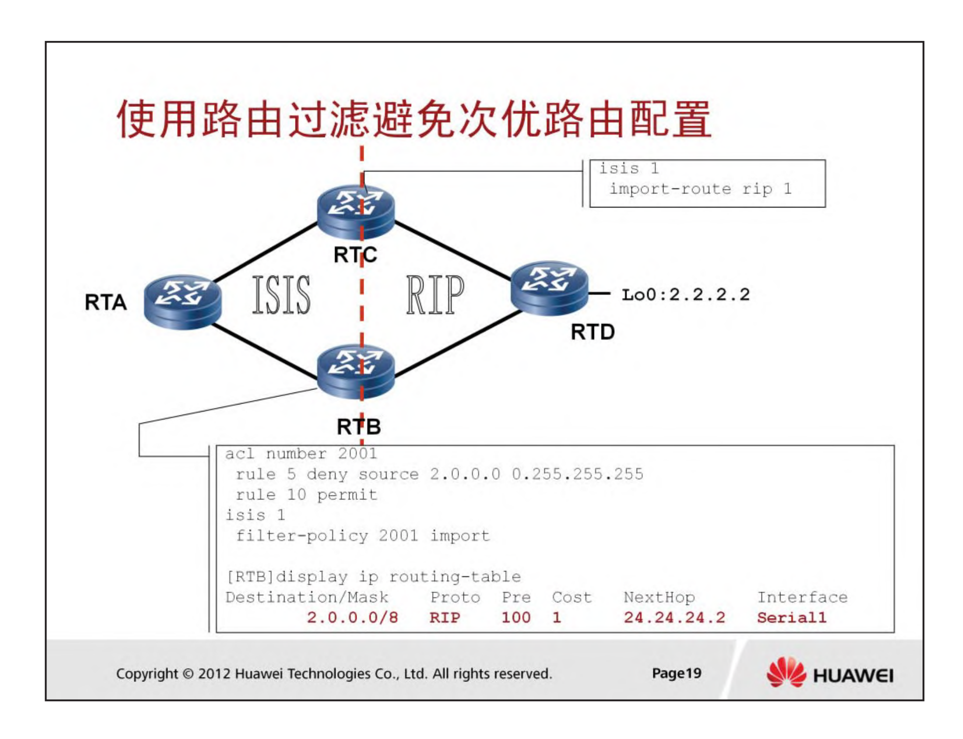


前面我们提到，路由引入可能会导致次优路由和路由环路，那么，我们可以采用路由过滤来避免这些问题。另外，路由过滤还可以使我们进行精确的路由引入和路由通告。

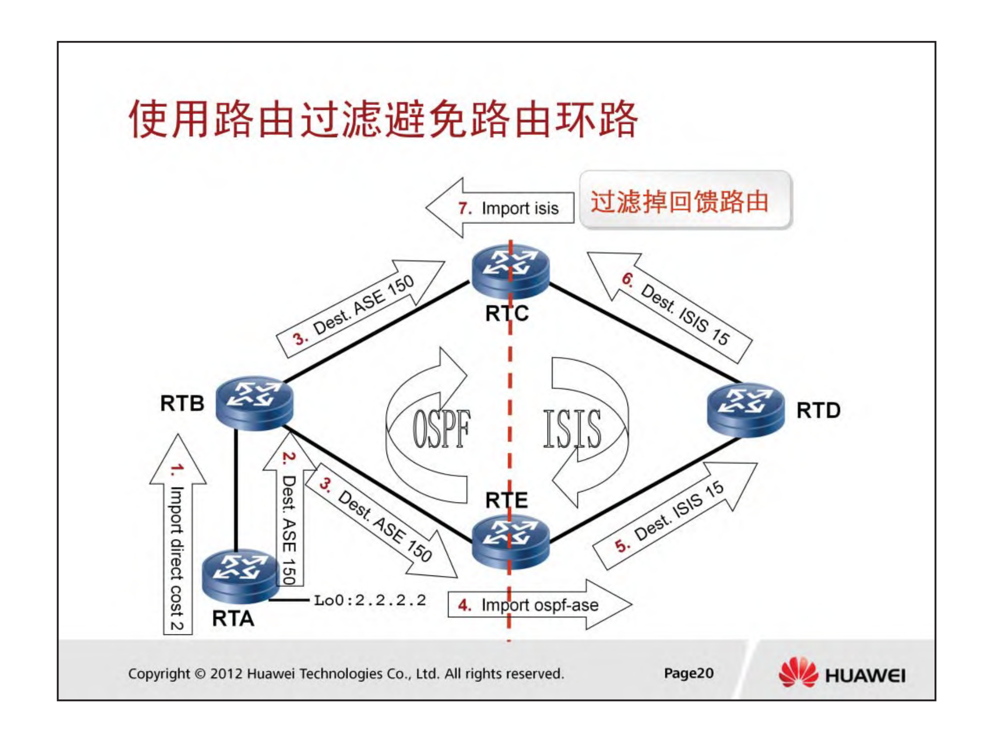


在这个网络中，同时运行ISIS和RIP协议。其中RTC和RTB是ASBR。在RTC上，我们把RIP引入到ISIS中。2.2.2.2这个网段将通过ISIS通告给RTA，然后RTA又通告给RTB。RTB同时从RIP和ISIS学习到关于2.2.2.2的路由，于是它比较ISIS和RIP的优先级，因为ISIS的优先级为15，RIP的优先级为100，所以RTB最终选择ISIS通告的路由。于是，RTB如果要把数据包发往2.2.2.2的话，将选用RTB-RTA-RTC-RTD的次优路径。

在这里，次优路由产生的原因是RTB同时从ISIS和RIP学到关于2.2.2.2的路由，而且RTB选择了从ISIS学到的路由。我们可以在RTB上配置路由过滤，把ISIS的路由过滤掉，这样，RTB将选择RIP路由来转发数据包，从而避免了次优路由。

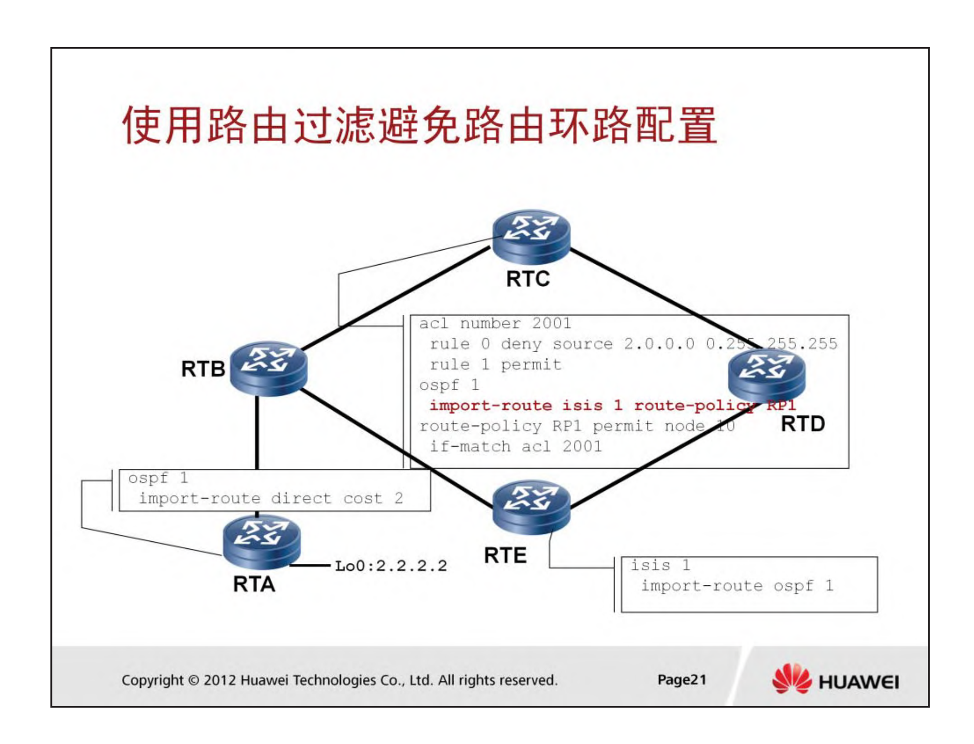


在RTB上配置filter-policy，把通过IS-IS获得的2.0.0.0的路由过滤掉，使得通过RIP获得的2.0.0.0路由出现在路由表中。

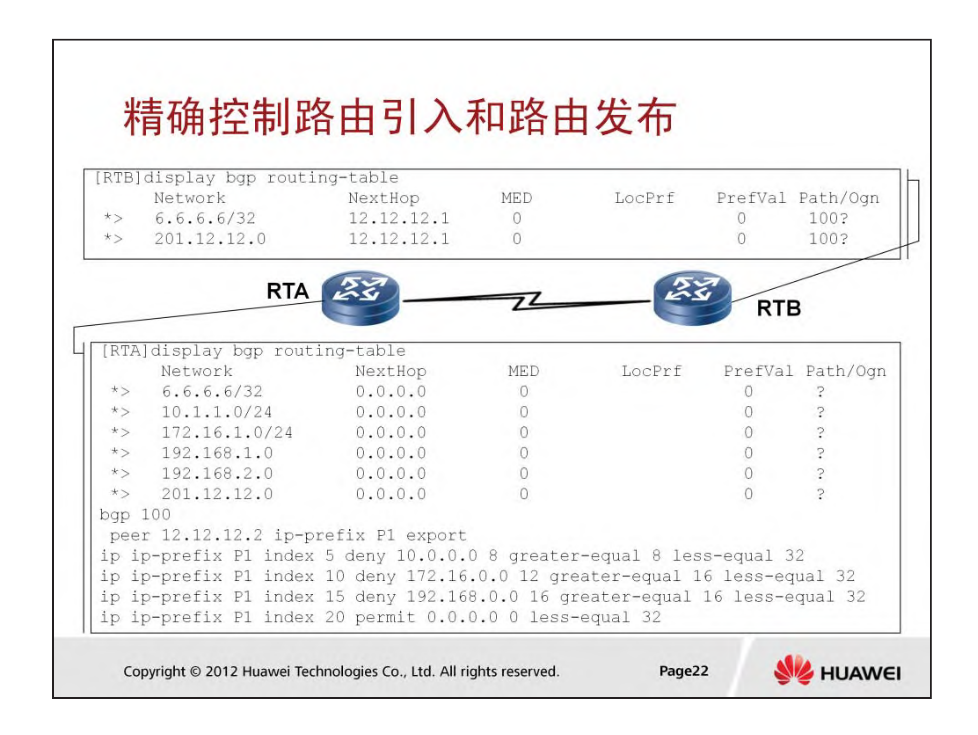


RTA通过引入直连路由把2.2.2.2网段引入到OSPF中，同时把引入路由的开销值设为2，OSPF将采用ASE路由（优先级为150）的方式把2.2.2.2通告给RTB, RTC, RTE。在RTE上，配置了引入OSPF-ASE，把2.2.2.2引入到ISIS中。在RTC上，配置了引入ISIS，把ISIS路由引入到OSPF中，于是，2.2.2.2网段又从ISIS通告回OSPF，这叫做路由回馈。这样的话，RTB同时从RTA和RTC学习到关于2.2.2.2的路由，因为优先级都一样（都是OSPF ASE路由），所以比较metric值，如果RTB很不幸地选择了RTC通告的路由，环路就产生了。比如：RTD发一个数据包到2.2.2.2，数据包将发往RTE，然后到RTB，因为RTB选择了RTC的路由，所以RTB把数据包发往RTC，RTC再发到RTD，于是数据包回到了起点。

在这里，路由环路产生的原因是路由回馈。所以，只要我们在RTC上配置路由引入的时候把2.2.2.2的路由过滤掉就可以了。



在RTC上把ISIS路由引入到OSPF时使用路由策略，把2.0.0.0路由过滤掉，从而避免环路。



在通告路由的时候，我们不希望把私网路由通告到公网中去，也可能需要隐藏内部网络的某些路由信息。我们可以使用路由过滤来精确控制路由信息的发布。

在引入路由的时候，我们可能不希望引入所有路由，只希望引入某些特殊的路由，我们可以使用路由过滤来精确控制路由引入。

在这个例子中，路由表里包括三类私有路由，我们需要定义前缀列表过滤掉私有路由。“ip ip-prefix P1 index 5 deny 10.0.0.0 8 greater-equal 8 less-equal 32”过滤掉10.0.0.0~10.255.255.255的私有路由；“ip ip-prefix P1 index 10 deny 172.16.0.0 12 greater-equal 16 less-equal 32”过滤掉172.16.0.0~172.31.255.255的路由；“ip ip-prefix P1 index 15 deny 192.168.0.0 16 greater-equal 16 less-equal 32”过滤掉192.168.0.0~192.168.255.255的路由；“ip ip-prefix P1 index 20 permit 0.0.0.0 0 less-equal 32”允许其它路由通过。

路由过滤规则

可以在出方向过滤路由

- 只能过滤路由信息，链路状态信息是不能被过滤的

可以在入方向过滤路由

- 对于链路状态路由协议，仅仅是不把路由加入到路由表中

可以过滤从其它路由协议引入的路由

- 只能在出方向过滤
- 在入方向过滤是没有意义的

可以使用filter-policy进行过滤，也可以使用ip-prefix进行过滤

 Page23 

路由过滤只能过滤路由信息，链路状态信息是不能被过滤的。对OSPF来说，只能过滤3类、5类、7类路由。过滤的方向可以是出方向和入方向。对于链路状态路由协议，如OSPF和ISIS，在入方向过滤路由实际上并不能阻断链路状态信息的传递，过滤的效果仅仅是路由不能被加到本地路由表中，而它的邻居仍然可以收到完整的路由状态信息并计算出完整的路由。

路由过滤还可以针对从其它协议引入的路由进行过滤，比如，把RIP路由引入到OSPF，OSPF可以使用路由过滤把某些从RIP引入的路由过滤掉，不通告给其它邻居。这种配置只能用在出方向上。



目　录

Page24



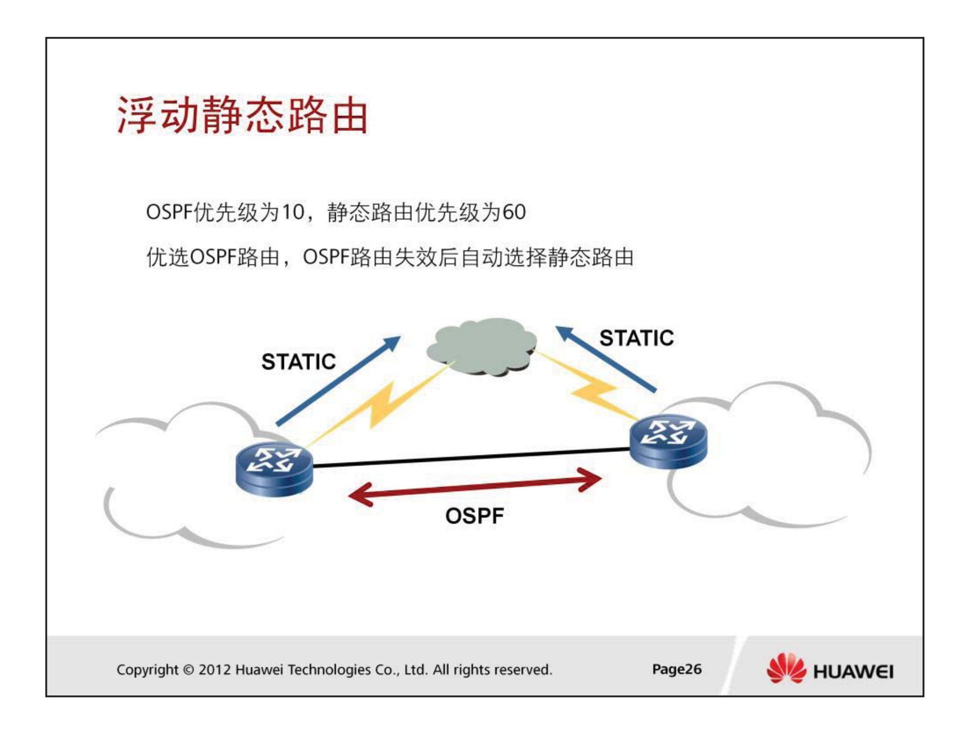
路由协议优先级

路由协议或路由种类	优先级
Direct	0
OSPF	10
IS-IS	15
Static	60
RIP	100
OSPF ASE	150
BGP	255

 Page25

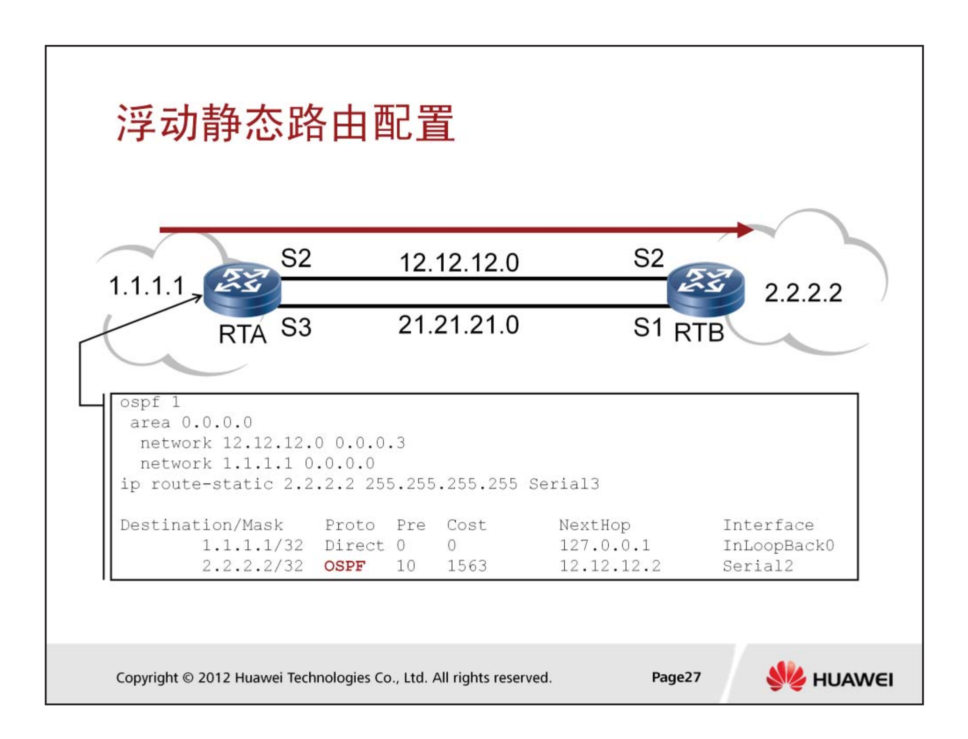


不同厂商定义的路由协议的优先级是不一样的。路由协议优先级的作用是给不同协议发现的路由分配不同的优先级，这样当一个路由器同时从不同的路由协议学习到相同的路由时，可以有一个选择的优先顺序。

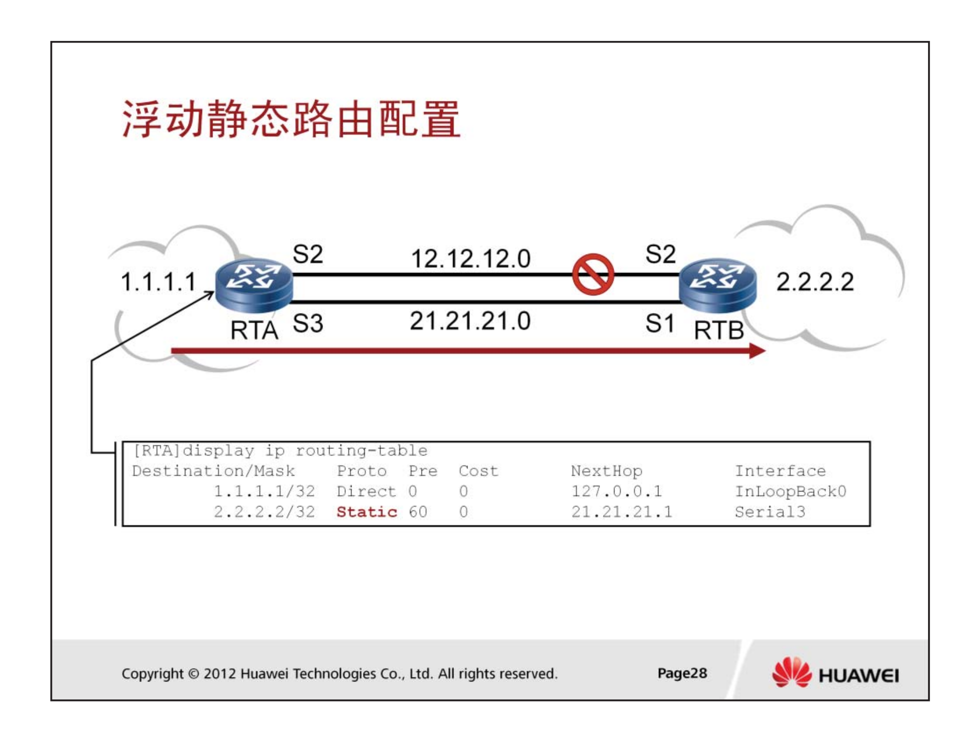


浮动静态路由是路由协议优先级的一个典型用途。在很多情况下，我们使用主备链路的方式来连接远程网络。在主链路上，我们一般运行动态路由协议，比如OSPF，ISIS等。备用链路一般是拨号链路，费用昂贵，而且是按连接时长收费的，所以，在拨号链路上通常不运行动态路由协议，只是配置一条指向远端网络的静态路由。在主链路正常的情况下，路由器可以从OSPF和静态配置学习到相同的路由，因为我们配置静态路由的优先级比OSPF路由的优先级低，所以路由器选择OSPF学习到的路由，数据包通过主链路进行转发。当主链路出现故障，OSPF邻居中断，从OSPF学习到的路由也随之失效并从路由表中清除。这个时候，原来配置的静态路由就自动“浮”出来，加入到路由表中，于是数据包就可以沿着备用链路进行转发了。当主链路恢复正常，OSPF邻居又重新建立，于是OSPF路由重新取代静态路由，流量又切换到主链路上，备用链路也自动DOWN掉。

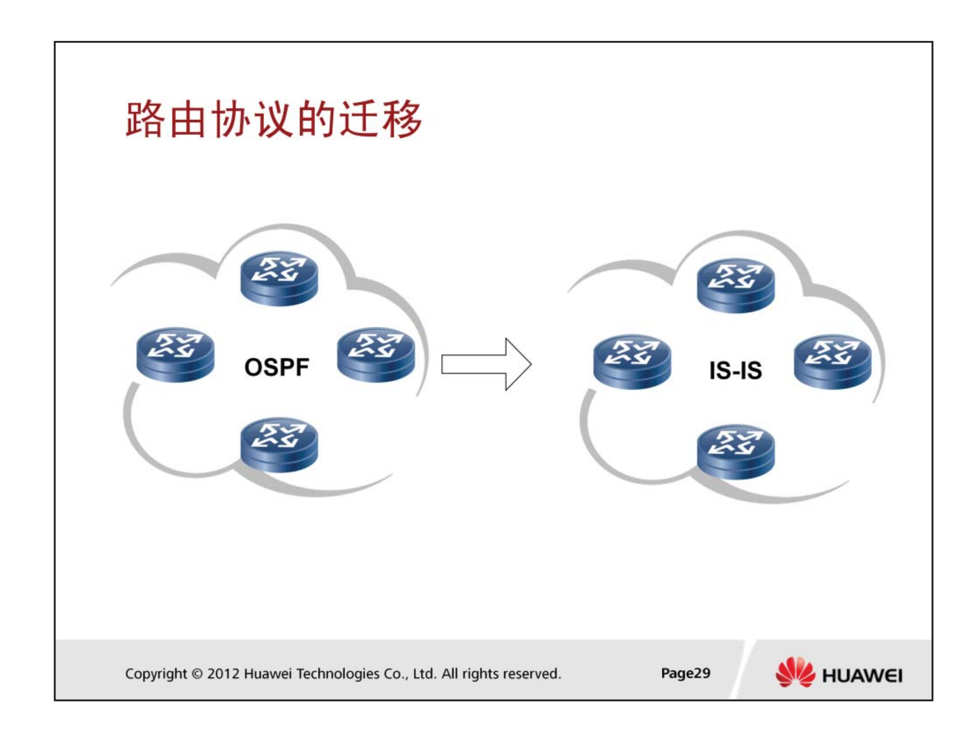
这种配置方式既节约成本，又增加了网络的可靠性，但是这种方法不能实现负载分担。



在这个示例中，RTA有两条链路连接到RTB，其中通过S2接口的链路和RTB通过OSPF学习到2.2.2.2网段的路由，另外还配置静态路由指向2.2.2.2，静态路由的出口是S3。在两条链路都正常的情况下，RTA选择OSPF路由到达2.2.2.2，在路由表中，我们可以清楚看到目的网段为2.2.2.2的路由是通过OSPF学习到的。



如果上面的链路发生了故障，RTA和RTB将不能交互hello报文，导致邻居关系失效，邻居关系失效会导致RTA路由表中从OSPF学到的2.2.2.2网段的路由失效，于是原来配置的静态路由将会出现在路由表中，这时RTA将通过S3接口访问2.2.2.2网段。



有的时候，我们需要使用新的路由协议来取代原有网络中的路由协议，而且要尽可能减小协议迁移时的网络中断时间。

比如原来网络使用OSPF协议，现在要迁移到ISIS协议。我们可以在每台路由器上同时运行两种路由协议，适当调整两种路由协议的优先级，使得在最初的时候IS-IS只在后台运行，在对IS-IS的邻居关系以及LSDB等参数经过仔细检查之后，对IS-IS的优先级进行修改从而使得IS-IS取代当前运行的路由协议。

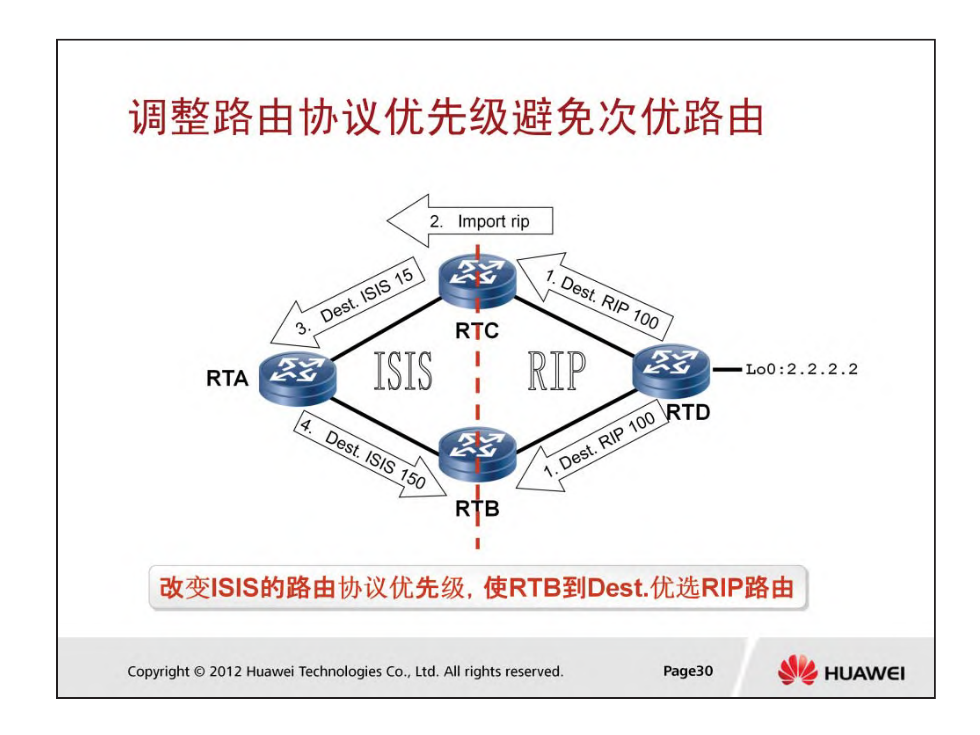
方法步骤：

1、此时网络中只运行OSPF，执行硬件和软件检查并进行升级以确保硬件和软件能够支持迁移；

2、配置IS-IS，设置合适的优先级，保证IS-IS只在后台运行，这样在每台路由器上都建立了IS-IS链路状态数据库，但路由表和转发表保持不变。在这个阶段对IS-IS的运行情况进行验证，只有当每台路由器都建立了LSDB并且数据库中已存在所有预期的LSP，同时已确认可以产生反映当前IP路由表中路由的IS-IS路由时，这个阶段可以结束；

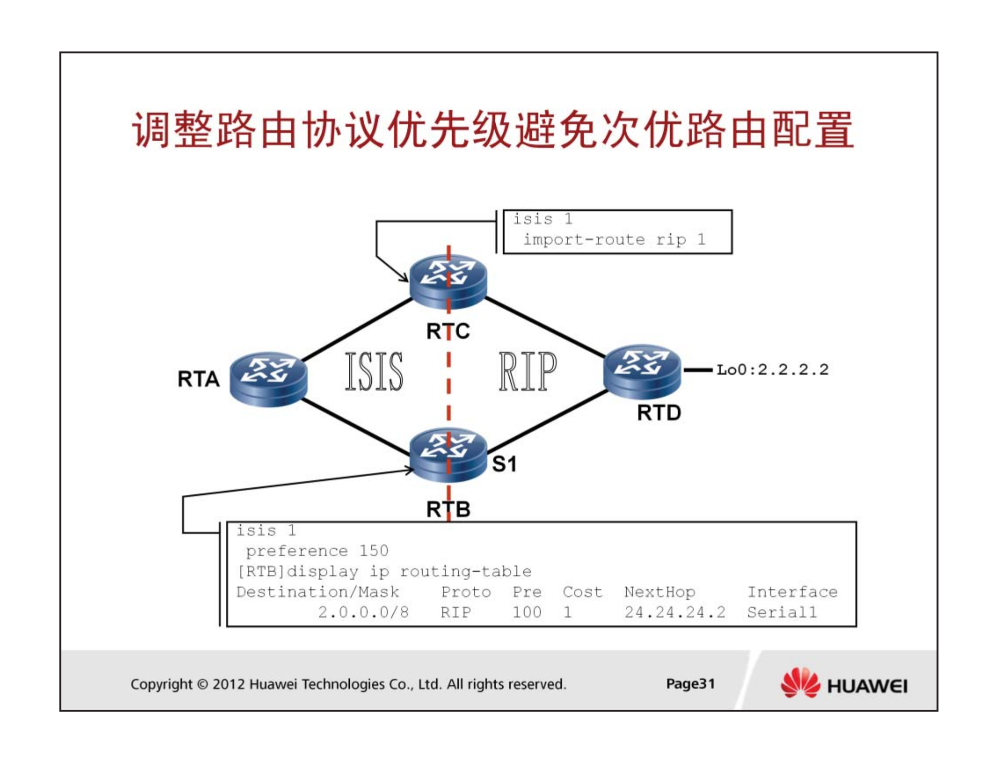
3、改变优先级，使得IS-IS取代原来的IGP，从而在前台运行；

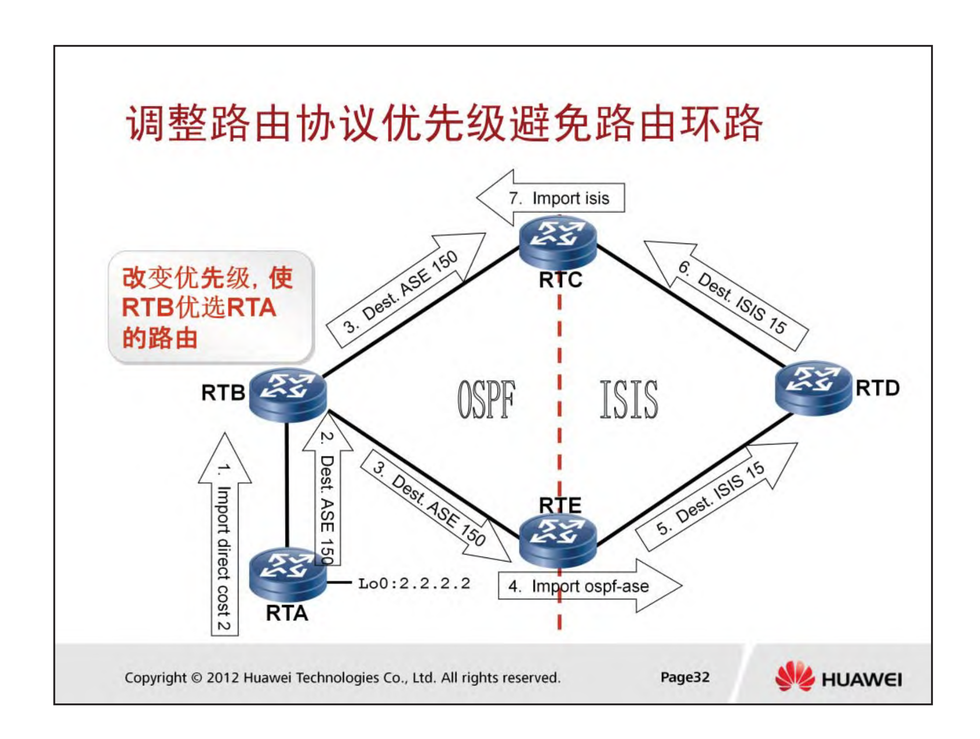
4、网络运行正常后，把原来的IGP删除。



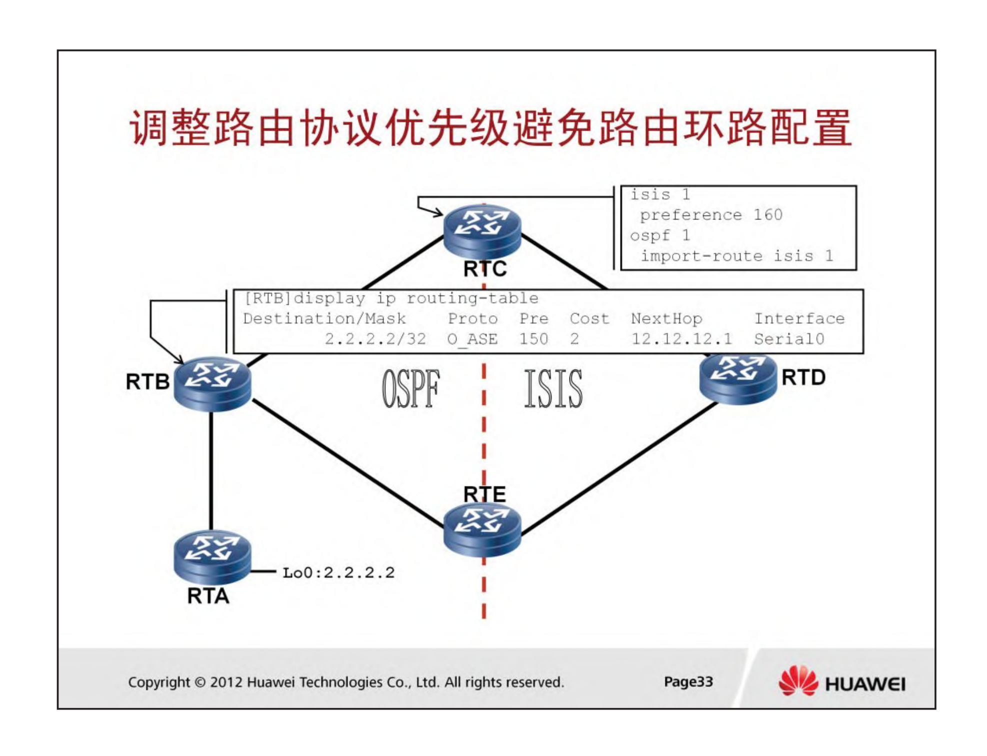
合理调整路由协议的优先级可以避免次优路由的产生。在这个拓扑图中，我们在RTB上改变ISIS的优先级，这样RTB会优选RIP的路由从而避免次优路由。

在调整路由协议优先级的时候一定要小心，避免出现新的问题，导致路由混乱。


调整路由协议优先级避免次优路由配置
isis 1
import-route rip 1
RTC
RTA
ISIS
RIP
Lo0:2.2.2.2
RTD
S1
RTB
isis 1
preference 150
[RTB]display ip routing-table
Destination/Mask Proto Pre Cost NextHop Interface
2.0.0.0/8 RIP 100 1 24.24.24.2 Serial1
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page31
HUAWEI



如果我们能够改变某条路由的路由优先级，还可以实现避免环路。图中，如果在RTC上把ISIS的路由优先级设为160，就不会有环路发生了。


调整路由协议优先级避免路由环路配置
isis 1
preference 160
ospf 1
import-route isis 1
RTC
[RTB]display ip routing-table
Destination/Mask Proto Pre Cost NextHop Interface
2.2.2.2/32 O_ASE 150 2 12.12.12.1 Serial0
RTB
OSPF
ISIS
RTD
RTE
Lo0:2.2.2.2
RTA
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page33
HUAWEI





目 录

路由引入

路由过滤

使用路由协议优先级

控制缺省路由下发

 Page34



路由协议发布缺省路由

所有路由协议都可以下发缺省路由

OSPF可以配置多种下发方式

- 在ABR上下发
- 在ASBR上下发
- 强制下发
- 非强制下发

IS-IS没有命令实现类似OSPF的非强制下发功能，但是可以通过路由策略来实现

 Page35



OSPF下发缺省路由比较复杂。根据区域类型的不同，下发的情况不一样。根据配置位置的不同，下发的情况也不一样。

按照区域来讨论，OSPF的区域类型包括普通区域、STUB区域、完全STUB区域、NSSA区域。

当区域被缺省定义时，它被认为是普通区域。普通区域可以是标准区域或骨干区域。标准区域是最通用的区域，它携带区域内路由，区域间路由和外部路由。骨干区域是连接所有其它OSPF区域的中央区域。

在普通区域里，默认是不产生缺省路由的。可以在ASBR上配置强制下发缺省路由或非强制下发TYPE5类型的缺省路由。

如果配置非强制下发，只有当路由表中存在一条活动缺省路由，而且这条缺省路由不是本OSPF进程产生的情况下，才下发缺省路由。同时，路由器还会学习其它OSPF路由器下发的TYPE5缺省路由，如果其它OSPF路由器下发的缺省路由优于本地路由表中的活动缺省路由，则路由器使用其它OSPF路由器下发的缺省路由来取代本地路由表中的活动缺省路由，并且停止下发TYPE5缺省路由。

如果配置强制下发，则不管路由表里是否存在缺省路由都会下发。同时，路由器不会学习其它OSPF路由器下发的TYPE5缺省路由。

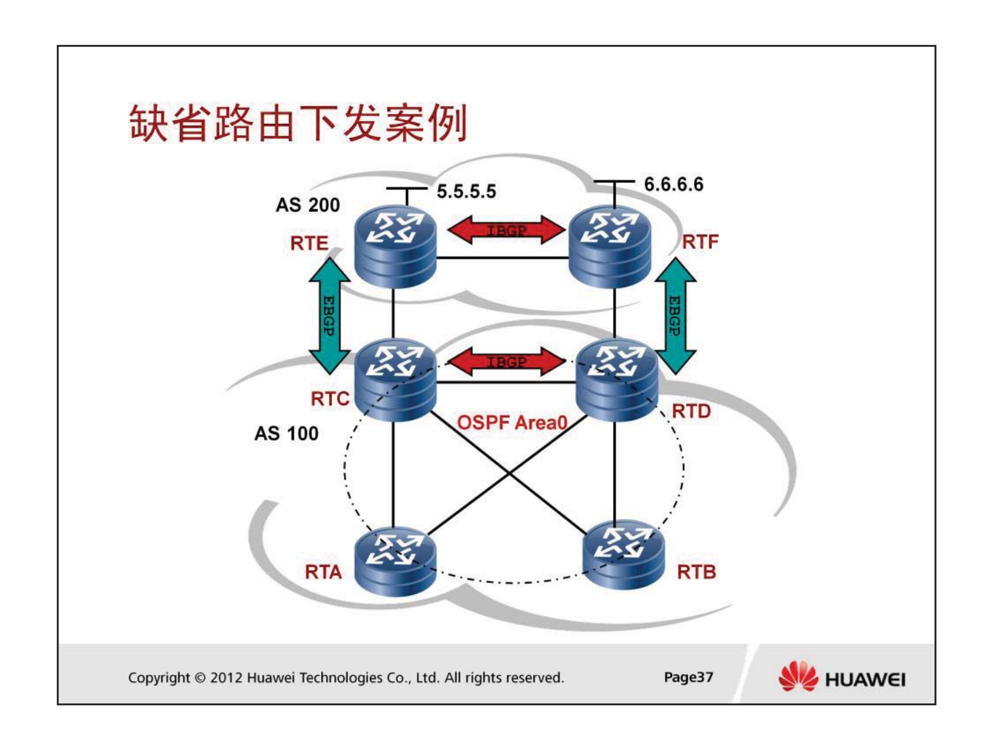
TYPE5的LSA会在整个路由域泛洪。

OSPF下发缺省路由

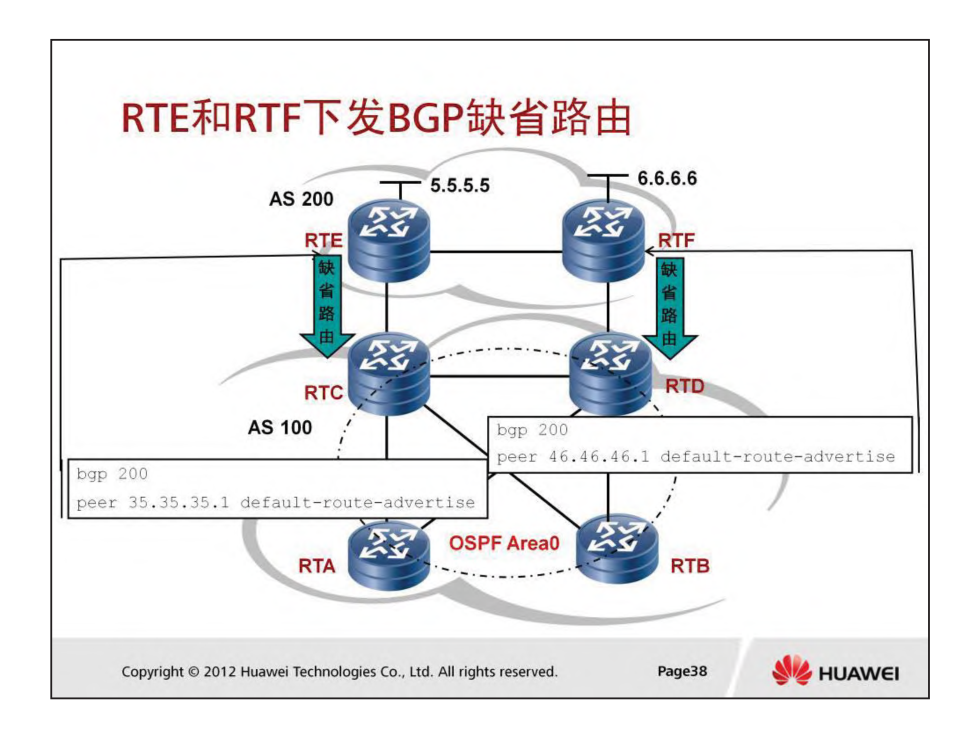
区域类型	自动产生	配置命令	产生者	已存在缺省路由	LSA类型	范围
普通		default-route-advertise	ASBR	√	5	路由域
		default-route-advertise always	ASBR		5	路由域
STUB	√		ABR		3	STUB
完全STUB	√		ABR		3	STUB
NSSA		nssa default-route-advertise	ABR		7	NSSA
		nssa default-route-advertise	ASBR	√	7	NSSA

HUAWEI

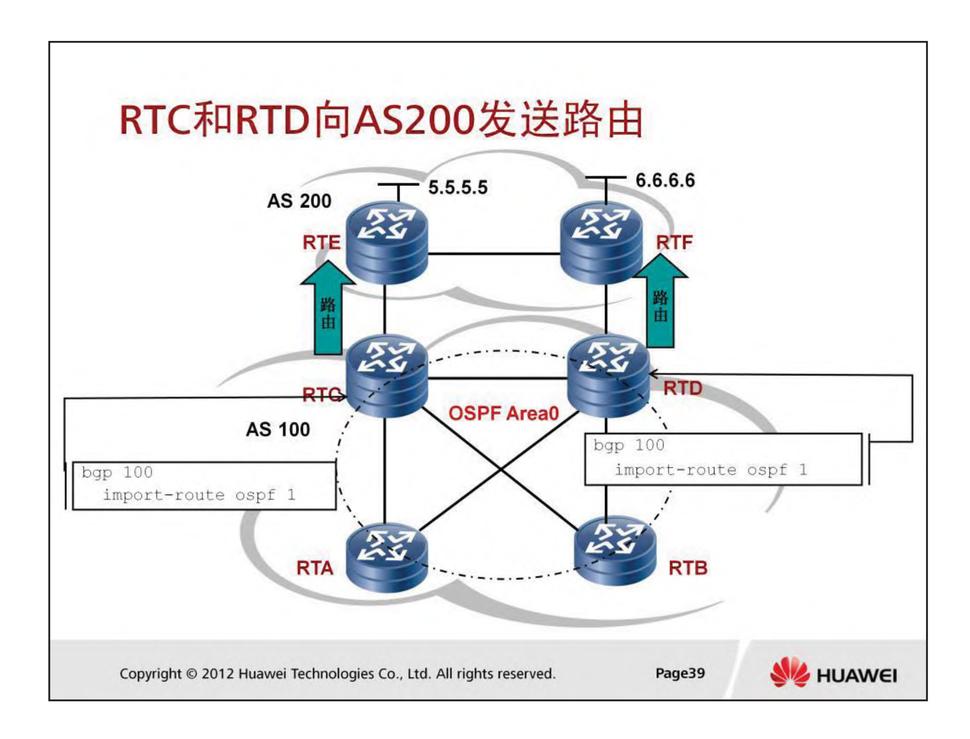
OSPF下发缺省路由的方式比较复杂，具体参加上表。



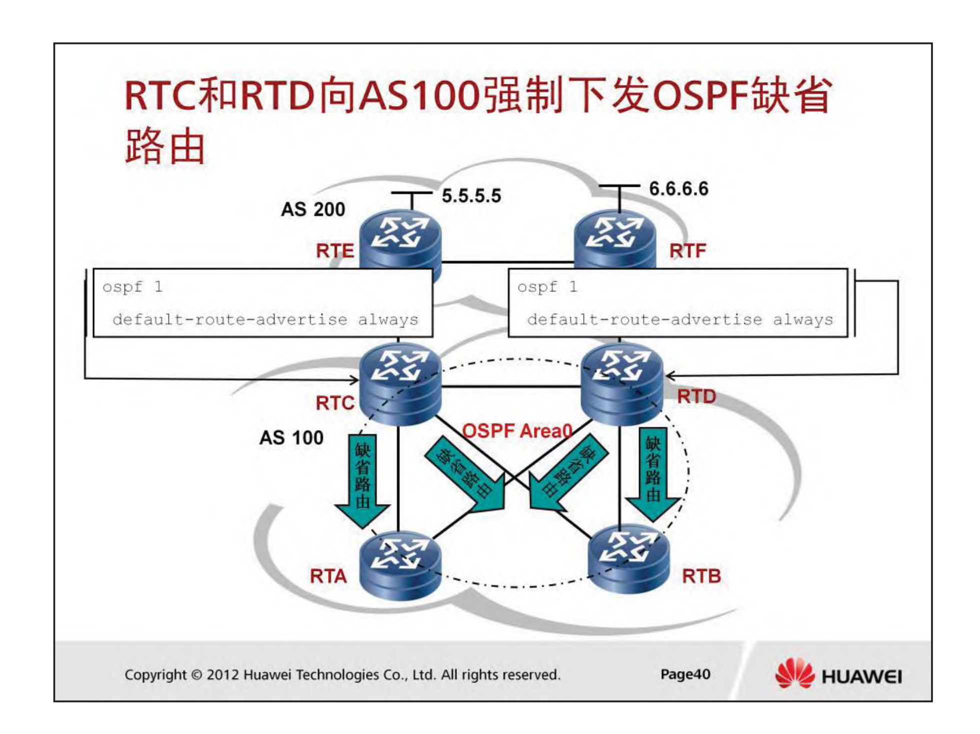
在现实的组网中，经常可以见到图中的这种组网方式。AS100有两个出口路由器RTC和RTD上行到AS200，其中RTE和RTC之间以及RTF和RTD之间分别建立EBGP邻居关系，RTE和RTF之间以及RTC和RTD之间分别建立IBGP邻居关系。AS100内部运行OSPF路由协议。



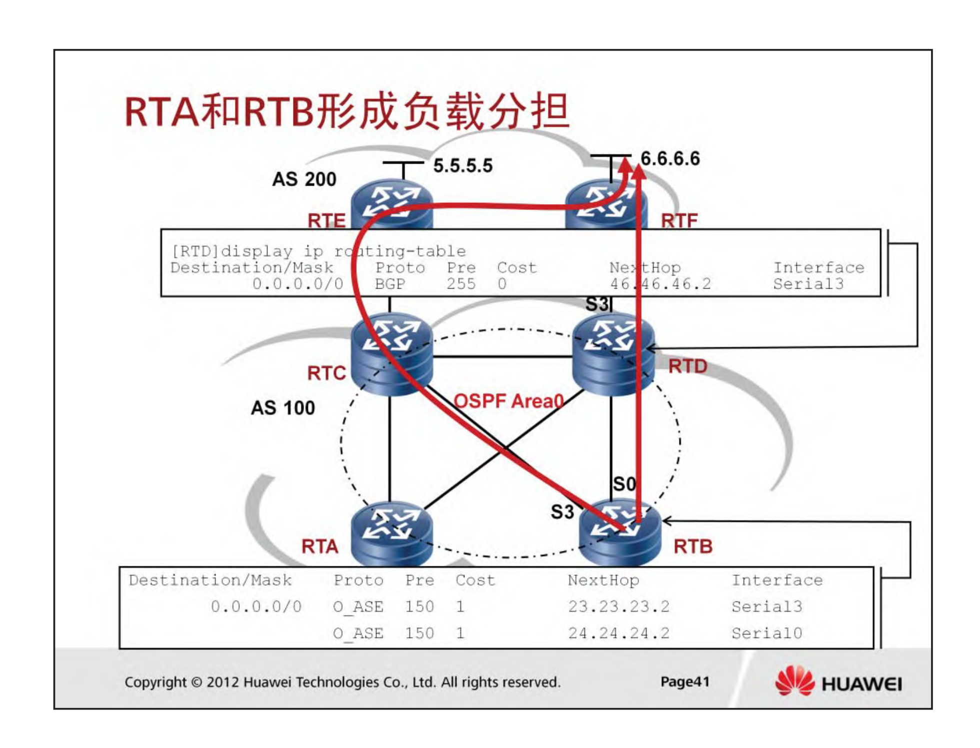
AS100不需要知道AS200的所有BGP路由，所以在RTE和RTF上下发BGP缺省路由给RTC和RTD。



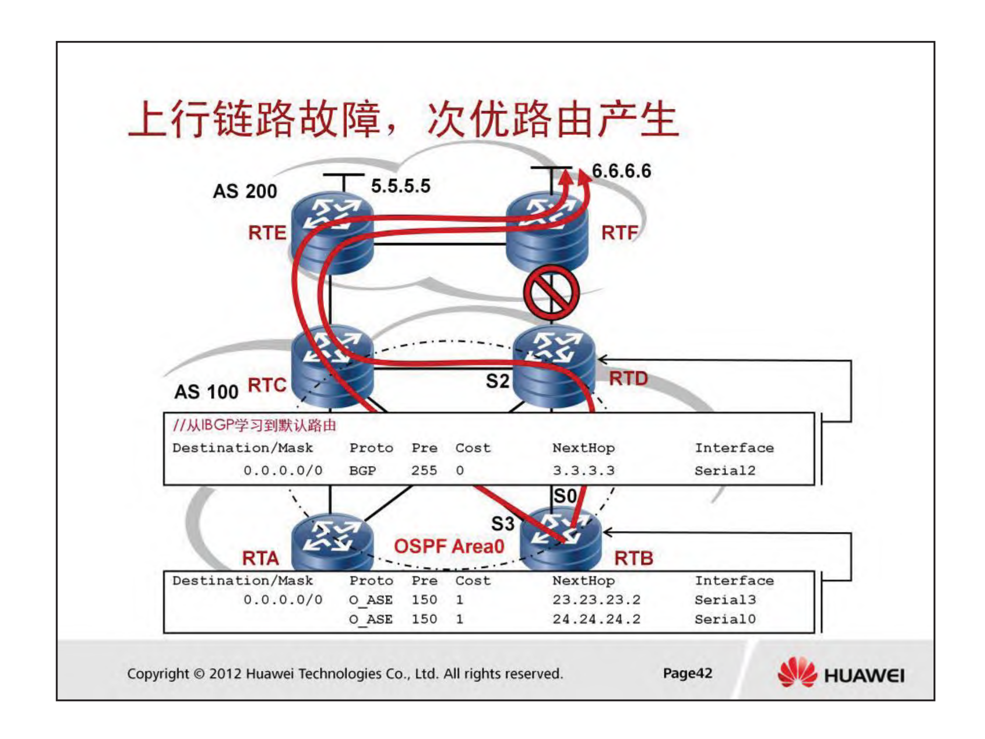
把OSPF路由引入到BGP中，这样AS200就可以把数据发到AS100。



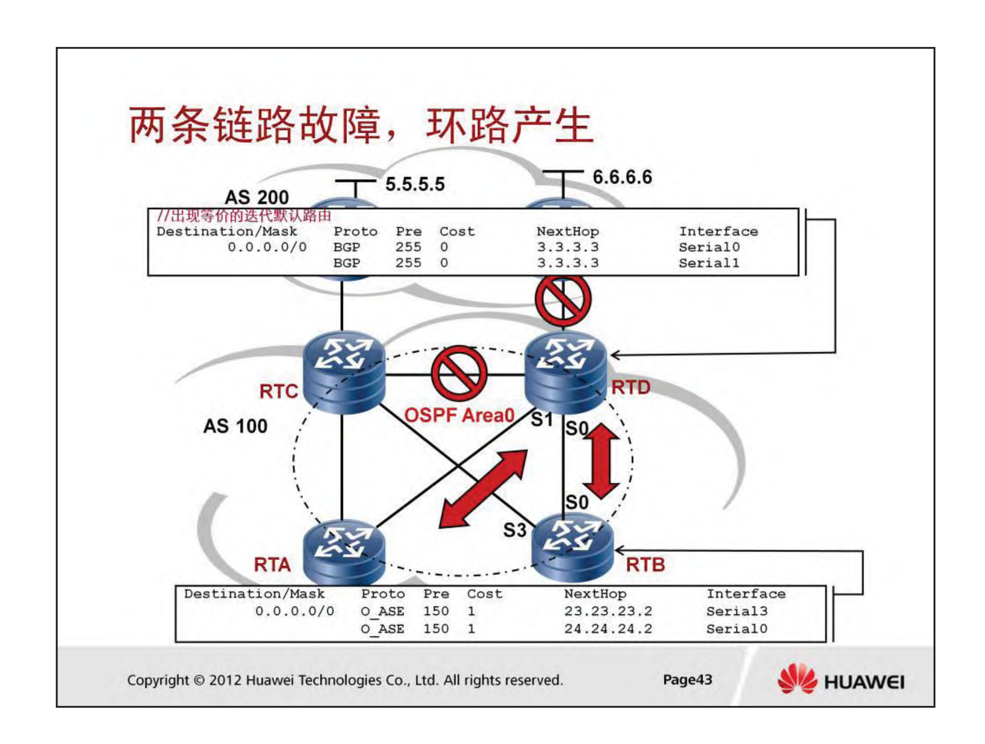
因为RTA和RTB并没有运行BGP协议，因此还不能访问AS200。在RTC和RTD上强制下发OSPF缺省路由，这样RTA和RTB可以使用缺省路由把数据包发到出口路由器RTC和RTD上，然后RTC和RTD再根据BGP缺省路由将数据包发到AS200。



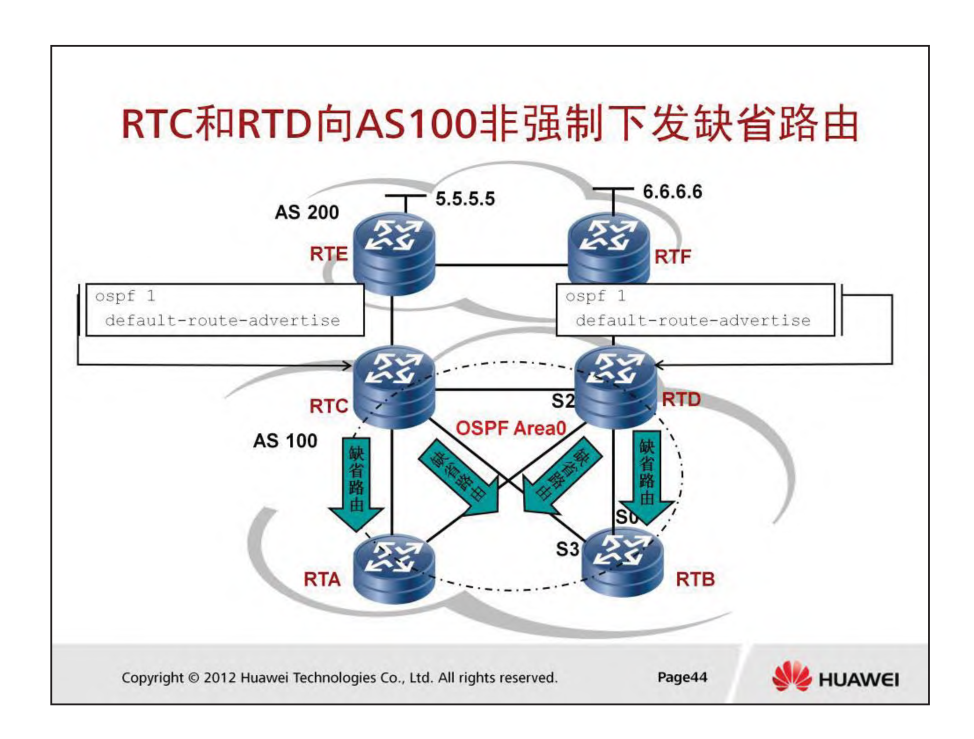
因为RTA和RTB同时从RTC和RTD学习到OSPF缺省路由，因此形成了负载分担，通过路由表可以证实这点。当RTA或者RTB要把数据包发到AS200，可能部分数据包选择RTC为出口，其它数据包选择RTD为出口。



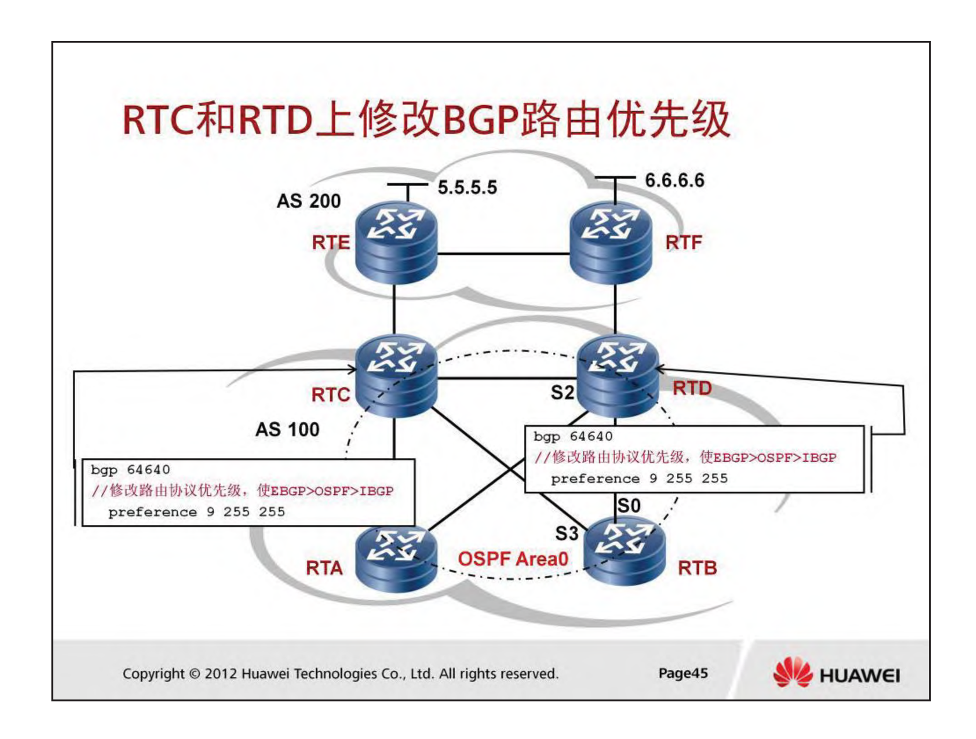
这种组网方式在链路一切正常的情况下可以工作得很好，但是一旦出现链路故障，就可能造成一些问题。在图中，RTD的上行链路出现了故障，RTD不能从RTF学习到BGP缺省路由，但是还可以从RTC学到缺省路由（在图中，RTD路由表的缺省路由下一跳是3.3.3.3，这是RTC的环回口地址），于是RTA和RTB仍然可以访问AS200，但是部分流量将通过RTC和RTD之间的链路，这是一个次优路由，而且如果设计不好的话还可能造成RTC和RTD之间的链路拥塞。



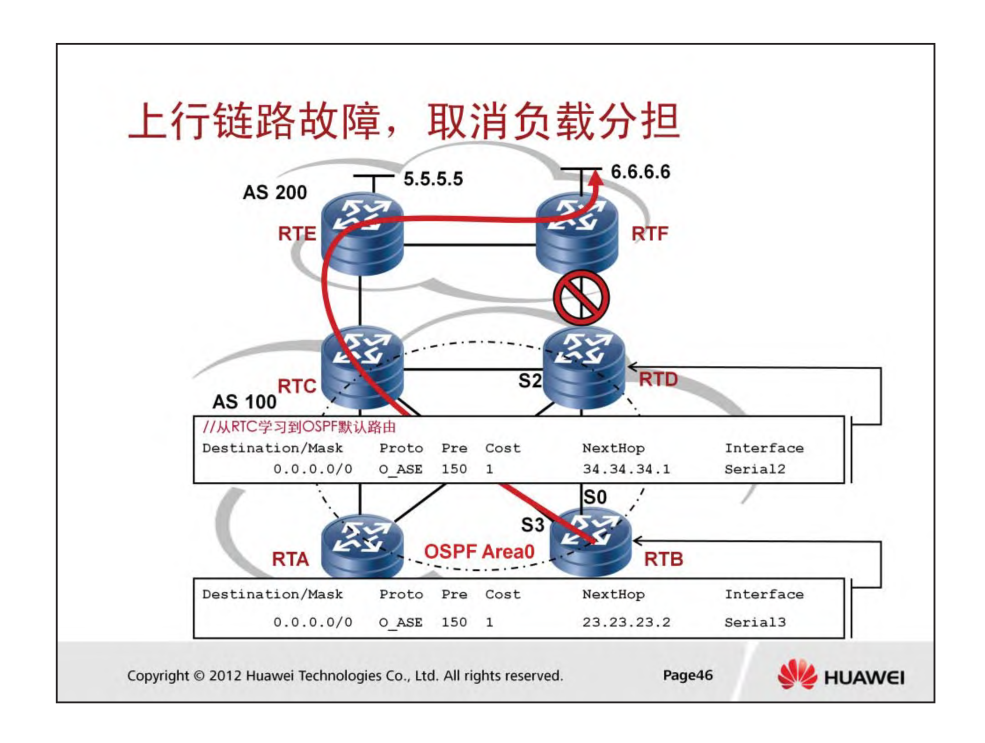
更严重的问题出现在RTD上行链路和RTC与RTD之间链路同时故障的时候。这时候，RTD不能从RTF收到EBGP的缺省路由，也不能通过RTC和RTD之间的链路收到RTC的IBGP缺省路由。但是RTC和RTD的IBGP邻居并没有失效，所以RTD仍然可以通过RTA和RTB学习到RTC通告的IBGP缺省路由，而且形成了负载分担。当RTB要把数据包发给AS200的时候，因为负载分担，一部分数据包发给RTC，然后RTC根据EBGP缺省路由把数据包正确转发出去；另一部分数据包发给RTD，这时候问题发生了，RTD查自己的路由表，发现了两条缺省路由，一条指向RTA，一条指向RTB，于是RTD也执行负载分担，再将其中一部分流量发给RTB，于是环路产生了。



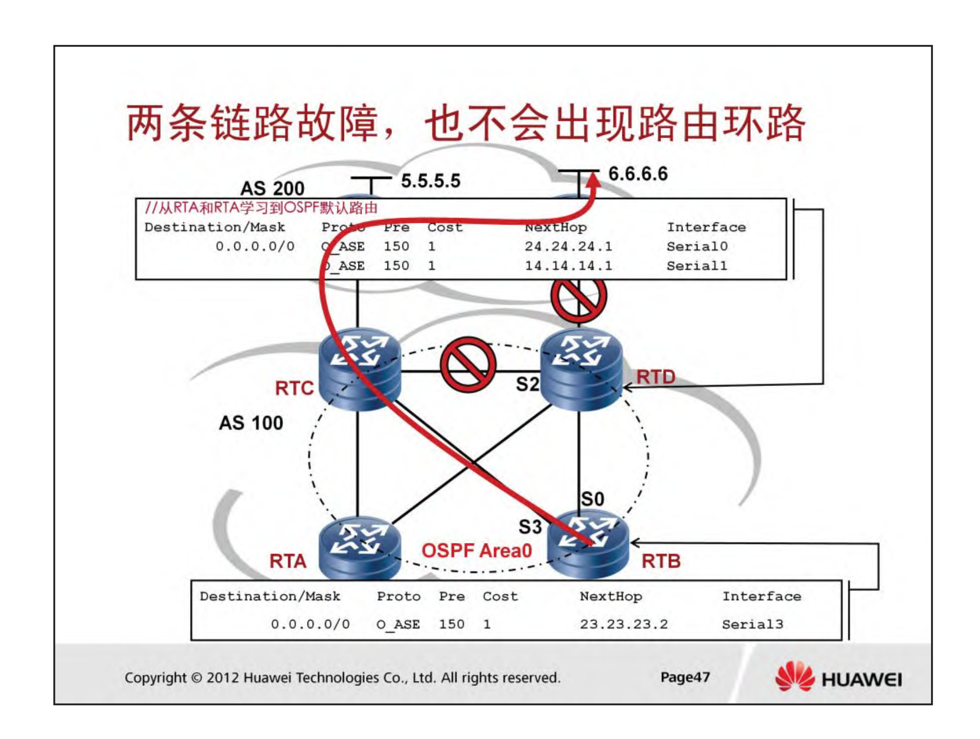
要解决前面出现的问题，需要在RTC和RTD上配置非强制下发缺省路由。



另外，在RTC和RTD上还要调整路由协议优先级，使EBGP路由优先级高于OSPF路由优先级，而OSPF路由优先级又高于IBGP路由优先级。



当RTD上行链路故障的时候，RTD不能再从RTF学习到EBGP缺省路由，RTD可以从RTC学习到IBGP缺省路由，但是RTD的路由表里出现的是OSPF的缺省路由（因为OSPF的协议优先级比IBGP协议优先级高）。根据非强制下发缺省路由的条件，RTD因为路由表里有从其它路由器学习到的OSPF缺省路由，所以RTD不再下发缺省路由，这时在RTA和RTB上只存在一条指向RTC的缺省路由。这时RTA和RTB发往AS200的数据包只会以RTC为出口。



在两条链路都出现故障的时候，同样道理，RTD也不会下发缺省路由，所以也不会出现问题。





Q：本课程介绍了哪些路由策略?

A：路由引入；路由过滤；调整路由协议优先级；默认路由下发。

Q：怎样避免次优路由?

A：次优路由的产生原因很多，要根据具体原因具体分析，本课程介绍了使用路由过滤和改变路由协议优先级的方法来避免次优路由。

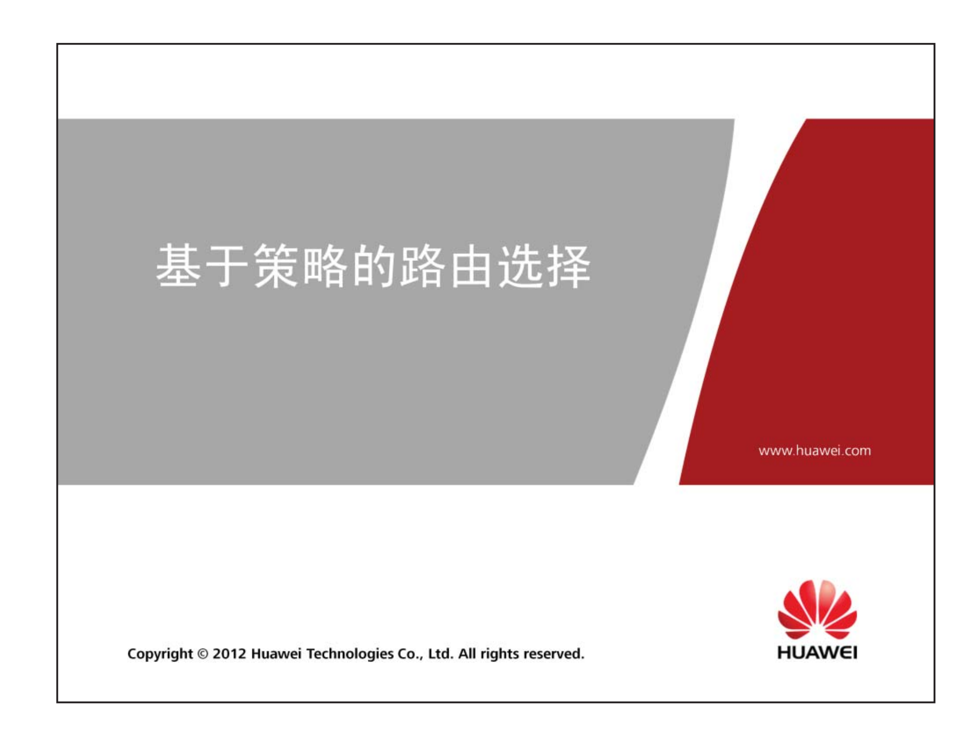
Q：怎样避免路由环路?

A：路由环路的产生原因很多，要根据具体原因具体分析，本课程介绍了使用路由过滤和改变路由协议优先级的方法来避免路由环路。

Q：非强制默认路由下发和强制默认路由下发有什么不同?

A：简单地说，非强制下发在满足一定条件（比如存在某条特定的路由）才下发缺省路由；强制下发则是无条件下发缺省路由。


谢谢
www.huawei.com


基于策略的路由选择
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

策略路由和路由策略都可以影响数据包的转发过程，但它们对数据包的影响方式是不同的。

Page1





培训目标

学完本课程后，您应该能：

- 理解策略路由和路由策略的区别
- 学会使用策略路由



基于策略的路由选择

策略路由是一种依据用户制定的策略进行报文转发路径选择的机制，与单纯依照IP报文的目的地址查找路由表进行转发不同，可应用于安全、QoS、负载分担等目的

策略路由支持基于ACL、报文长度等信息，来灵活地指定数据包的转发路径

Page3



策略路由的作用

为网络管理者提供了比传统路由协议对报文的转发和存储更强的控制能力

- 使网络管理者不仅能够根据目的地址，而且能够根据协议类型、报文大小、应用、IP源地址或者其它的策略来选择转发路径
- 可以控制多个路由器之间的负载分担、单一链路上报文转发的QoS或者满足某种特定需求

Page4



策略路由和路由策略的区别

策略路由可以不按照路由表进行报文的转发

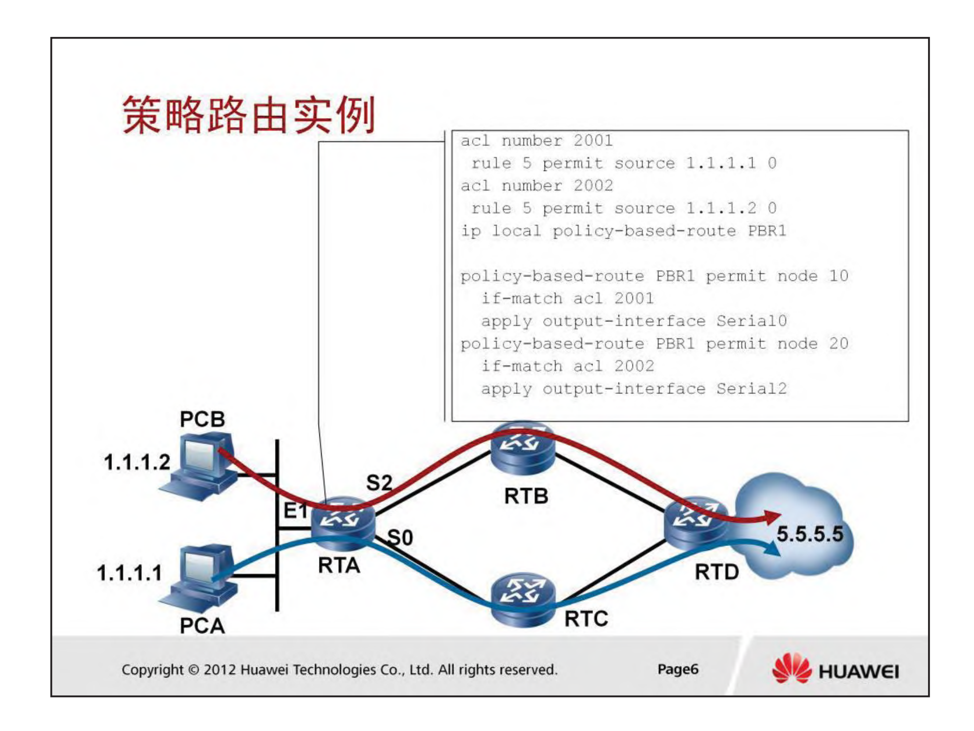
路由策略主要控制路由信息的引入、发布、接收等

Page5



策略路由主要是控制报文的转发，即可以不按照路由表进行报文的转发（因为一般报文的转发要通过查找转发表，而配上策略路由后就不用管转发表了，可以随心所欲将报文从转发出去了）。

路由策略主要控制路由信息的引入（控制哪些路由信息引到路由协议中，哪些路由信息不引入，主要是针对某种路由协议，是否允许其它路由信息引进来）、发布（控制哪些发布出去，哪些不发布出去，通过同一种路由协议发布出去）、接收（控制哪些接收，哪些丢弃）。



在这个拓扑中，如果没有负载分担的话，RTA的路由表中到达5.5.5.5的路由要么是通过RTB，要么是通过RTC。我们可以在RTA上配置策略路由，使得从1.1.1.1发出的数据包通过RTC到达5.5.5.5，而1.1.1.2发出的数据包通过RTB到达5.5.5.5。

定义策略路由的匹配规则

1、进入系统视图：system-view

2、创建策略或一个策略节点：policy-based-route policy-name { deny | permit } node node-id

3、设置IP报文长度匹配条件：if-match packet-length minimum-length maximum-length

4、设置IP地址匹配条件：if-match acl acl-number

由策略名称指定的策略可以包含若干策略点，策略点由顺序号node-id来指定，顺序号的值越小则优先级越高，相应策略优先执行。重复创建policy-based-route时，新的配置将覆盖旧的配置。策略的具体内容由if-match和apply子句来指定。

permit表示对满足匹配条件的报文运用策略，deny表示对满足匹配条件的报文不运用策略。IP单播策略路由提供两种定义报文的方法：根据报文长度匹配和根据ACL规则匹配。一条策略中可以包含多条if-match子句，组合使用。

定义策略路由的动作

1、进入系统视图：system-view

2、创建策略或一个策略节点：policy-based-route policy-name { deny | permit } node node-id

3、设置报文优先级：apply ip-precedence precedence

4、指定报文的缺省下一跳：apply ip-address default next-hop ip-address1 [ip-address2]

5、指定报文的缺省出接口：apply default output-interface interface-type1 interface-number1 [interface-type2 interface-number2]

6、设置报文的下一跳：apply ip-address next-hop ip-address

7、指定报文的出接口：apply output-interface interface-type1 interface-number1 [interface-type2 interface-number2]

8、设置访问VPN实例：apply access-vpn vpn-instance vpn-instance-name&<1-6>

apply子句用于对满足匹配条件的报文进行设置，使报文能够按照指定的路径转发。一条策略中可以包含多条apply子句，组合使用。

可以指定多个下一跳或设置多个出接口，这种情况下，报文转发将采用负载分担的方式进行。流量只在多个下一跳之间进行负载分担，或者只在多个出接口之间进行负载分担。如果同时配置了出接口和下一跳，仅在出接口之间进行负载分担。

说明：出接口不能为以太接口等广播型接口。

应用策略路由

在系统视图下应用策略路由，此时的策略路由只对本地产生的报文起作用。

启动本地策略路由

1、进入系统视图：system-view

2、使能本地策略路由：ip local policy-based-route policy-name

本地策略路由只对本地产生的报文有效。只能配置一条本地策略。



问 题

策略路由和路由策略有什么区别?

策略路由的作用有哪些?

 Page8



Q：策略路由和路由策略有什么区别?

A：策略路由主要是控制报文的转发，即可以不按照路由表进行报文的转发（因为一般报文的转发要通过查找转发表，而配上策略路由后就不用管转发表了，可以随心所欲将报文从转发出去了）。

路由策略主要控制路由信息的引入（控制哪些路由信息引到路由协议中，哪些路由器不引入，主要是针对某种路由协议，是否允许其它路由信息引进来）、发布（控制哪些发布出去，哪些不发布出去，通过同一种路由协议发布出去）、接收（控制哪些接收，哪些丢弃）。

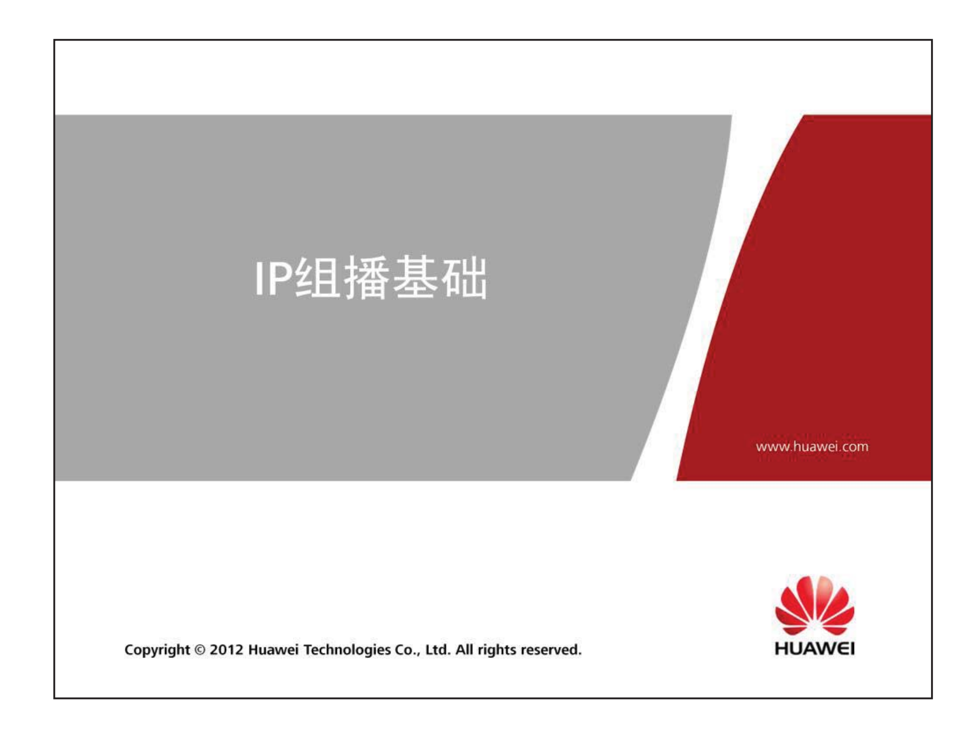
Q：策略路由的作用有哪些?

A：为网络管理者提供了比传统路由协议对报文的转发和存储更强的控制能力。使网络管理者不仅能够根据目的地址，而且能够根据协议类型、报文大小、应用、IP源地址或者其它的策略来选择转发路径。可以控制多个路由器之间的负载分担、单一链路上报文转发的QoS或者满足某种特定需求。


谢谢
www.huawei.com

Module 5

组播


IP组播基础
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

IP组播技术实现了IP网络中点到多点的高效数据传送，能够有效地节约网络带宽、降低网络负载，在实时数据传送、多媒体会议、数据拷贝、游戏和仿真等诸多方面都有广泛的应用。

 Page1



随着Internet网络的不断发展，网络中交互的各种数据、语音和视频信息越来越多，同时新兴的电子商务、网上会议、网上拍卖、视频点播、远程教学等服务也在逐渐兴起。这些服务对信息安全性、有偿性、网络带宽提出了要求。

现代网络传输技术对以下两项目标给予更高的关注：

资源发现

点对多点的IP传输

实现这两项目标有三种解决方案：单播（Unicast）、广播（Broadcast）、组播（Multicast）

本课程通过比较三种解决方案的数据传输方式，说明组播方式更适合点对多点的IP传输。



培训目标

学完本课程后，您应该能：

- 了解什么是组播及组播的地址结构
- 了解组播的转发流程
- 了解源路径树以及共享树等相关概念

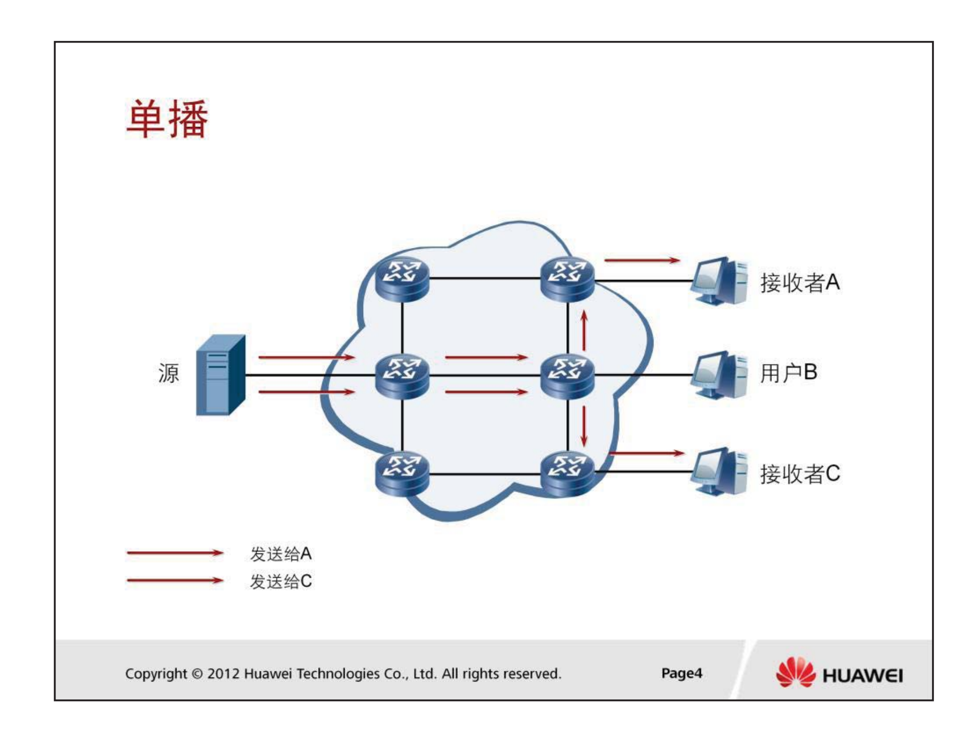
 Page2 

学习完此课程，您将会：

了解组播与单播以及广播三种数据传输方式的区别。掌握组播的地址结构以及组播报文的转发流程。掌握组播技术中的相关概念，如源路径树、共享树等。

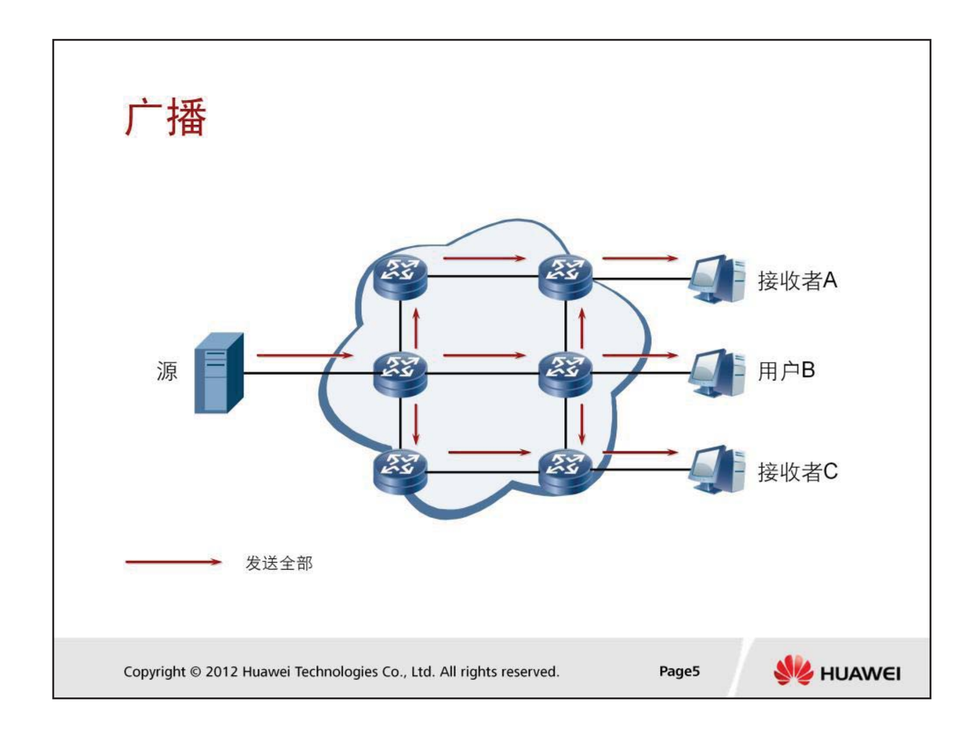


组播概述部分介绍组播与单播、广播三种数据传输方式的区别，并分析了组播技术的优劣势及具体应用。



网络中存在信息发送者“源”，接收者A和C提出信息需求，网络采用单播方式传输信息。

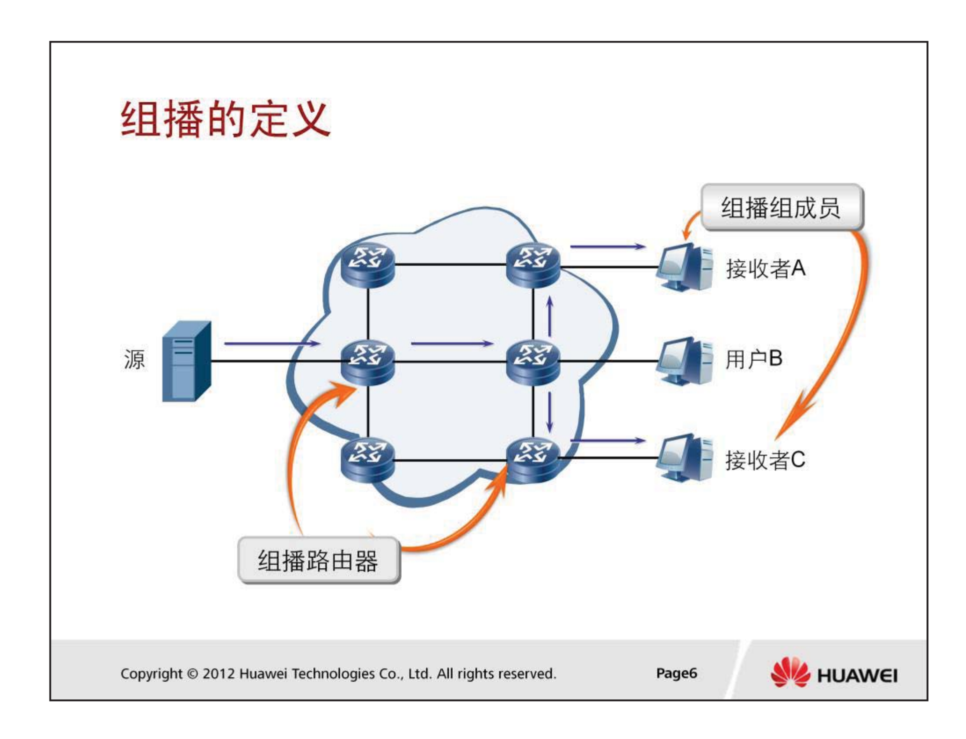
采用单播（Unicast）方式时，系统为每个需求该信息的用户单独建立一条数据传送通路，并为该用户发送一份独立的拷贝信息。由于网络中需求该信息的用户量和传输的信息量成正比，因此当需求该信息的用户量庞大时，网络中将出现多份相同信息流。此时，带宽将成为重要瓶颈，单播方式较适合用户稀少的网络，不利于信息规模化发送。



网络中同样存在信息发送者“源”，接收者A和C提出信息需求，网络采用广播方式传输信息。

广播（Broadcast）方式，系统把信息传送给网络中的所有用户，不管他们是否需要，任何用户都会接收到广播来的信息，信息安全性和有偿服务得不到保障。另外，当同一网络中需求该信息的用户量很小时，网络资源利用率将非常低，带宽浪费严重。

广播方式适合用户量很大的网络，当网络中需求某信息的用户量不确定时，单播和广播方式效率很低。



IP组播技术的出现及时解决了网络中用户数量不确定的问题。当网络中的某些用户需求特定信息时，组播信息发送者（即组播源）仅发送一次信息，借助组播路由协议为组播数据包建立树型路由，被传递的信息在尽可能远的分叉路口才开始复制和分发。

Multicast group称为组播组，使用一个IP组播地址标识。接收者A和C两个信息接收者，加入该组播组，从而可以接收发往该组播组的数据。

相比单播来说，使用组播方式传递信息，用户的增加不会显著增加网络的负载；不论接收者有多少，相同的组播数据流在每一条链路上最多仅有一份。使用组播方式传递信息，用户的增加不会显著增加网络的负载。相比广播来说，组播数据流仅会流到有接收者的地方，不会造成网络资源的浪费。

在组播方式中:

信息的发送者称为“组播源”。

接收相同信息的接收者构成一个组播组，并且每个接收者都是“组播组成员”。

提供组播功能的路由器称为“组播路由器”。

IP组播路由器不仅提供组播路由功能，也提供组成员管理功能。同时，自己本身也可以是一个或多个组播组的接收成员。同一组播组的成员

可以广泛分布在网络中的任何地方，即“组播组”关系没有地域限制。

组播优势和应用

组播的优势：

- 提高效率：降低网络流量、减轻硬件负荷
- 优化性能：减少冗余流量、节约网络带宽、降低网络负载。
- 分布式应用：使多点应用成为可能

组播的应用：

- 多媒体
- 培训、联合作业场合的通信
- 数据仓库、金融应用（股票）
- 任何的“单到多”数据发布应用

Page8



组播的优势主要在于：

提高效率：降低网络流量、减轻硬件负荷。

优化性能：减少冗余流量、节约网络带宽、降低网络负载。

分布式应用：使多点应用成为可能。

组播技术有效地解决了单点发送多点接收的问题，实现了IP网络中点到多点的高效数据传送。利用网络的组播特性可以方便地提供一些新的增值业务，包括在线直播、网络电视、远程教育、远程医疗、网络电台、实时视/音频会议等互联网的信息服务领域。

组播的劣势

组播是基于UDP的

- 尽力而为
- 没有拥塞避免机制
- 报文重复
- 报文失序

Page9


组播技术有效地解决了单点发送多点接收的问题，实现了IP网络中点到多点的高效数据传送。但由于组播技术是基于UDP的，所以同时也存在着不足之处:

尽力而为

报文丢失是不可避免的。因此组播应用程序不能依赖组播网络进行可靠性保证，必须针对组播网络的这个特点进行特别设计。“可靠组播”目前仍然处于研究阶段。

没有拥塞避免机制

缺少TCP窗口机制和慢启动机制，组播可能会出现拥塞。如果可能的话，组播应用程序应该尝试检测避免拥塞。

报文重复

某些组播协议的特殊机制（如Assert机制和SPT切换机制）可能会造成偶尔的数据包的重复。组播应用程序应该容忍这种现象。

报文失序

同样组播协议有的时候会造成报文到达的次序错乱，组播应用程序必须自己采用某种手段进行纠正（比如缓冲池机制等）。



本章介绍组播地址结构、地址分类以及组播MAC地址和单播MAC地址的对比。

组播IP地址

一个组播组就是一个IP地址，不表示具体的主机，而是表示一系列系统的集合，主机加入某个组播组即声明自己接收目的为某个IP地址的报文。

IP组播组地址

- 224.0.0.0–239.255.255.255
- “D”类地址空间
  - 第一个字节的高四位 = “1110”

 Page11



单播数据传输过程中，一个数据包传输的路径是从源地址路由到目的地址，利用“逐跳”（hop-by-hop）的原理在IP网络中传输。然而在IP组播环境中，数据包的目的地址不是一个，而是一组，形成组地址。所有的信息接收者都加入到一个组内，并且一旦加入之后，流向该组地址的数据立即开始向接收者发送，组中的所有成员都能接收到数据包，这个组就是“组播组”。

根据IANA（Internet Assigned Numbers Authority）规定，组播报文的目的地址使用D类IP地址，D类地址不能出现在IP报文的源IP地址字段中。

D类组播地址范围是从224.0.0.0到239.255.255.255。

组播IP模型分类

ASM（Any-Source Multicast）

SFM（Source-Filtered Multicast）

SSM（Source-Specific Multicast）

Page12



ASM模型

ASM模型就是任意源组播模型。在该模型中，任意发送者都可以成为组播源，向某组播组地址发送信息。众多接收者通过加入由该地址标识的主机组，从而接收到发往该组播组的所有信息。在ASM模型中，接收者无法预先知道组播源的位置，接收者可以在任意时间加入或离开该主机组。

SFM模型

SFM模型继承了ASM模型，从发送者角度来看，组播组成员关系完全相同。SFM在功能上对ASM进行了扩展：上层软件对接收到的组播报文的源地址进行检查，允许或禁止来自某些组播源的报文通过。最终，接收者只能接收到来自部分组播源的数据。从接收者角度来看，只有部分组播源是有效的，组播源经过了筛选。

SSM模型

在现实生活中，用户可能仅对某些源发送的组播信息感兴趣，而不愿接收其它源发送的信息。SSM模型为用户提供了一种能够在客户端指定信源的传输服务。

SSM模型和ASM模型的根本区别是接收者已经通过其他手段预先知道了组播源的具体位置。SSM使用和ASM不同的组播地址范围，直接在接收者和其指定的组播源之间建立专用的组播转发路径。

组播IP地址分类

永久组地址

- IANA为路由协议预留的组播地址，用于标识一组特定的网络设备（也称为保留组播组）。
  - 224.0.0.5 OSPF路由器
- 永久组地址保持不变，组成员的数量可以是任意的，甚至可以为零。

临时组地址

- 为用户组播组临时分配的IP地址，组成员的数量一旦为零，即取消。

 Page13 

永久组地址：IANA为路由协议预留的组播地址，用于标识一组特定的网络设备（也称为保留组播组）。永久组地址保持不变，组成员的数量可以是任意的，甚至可以为零。如224.0.0.5是为OSPF路由协议中预留的组播地址。

临时组地址：为用户组播组临时分配的IP地址，组成员的数量一旦为零，即取消。

组播IP地址分类

D类地址范围	含义
224.0.0.0~224.0.0.255	为路由协议预留的永久组地址。
224.0.1.0~231.255.255.255 233.0.0.0~238.255.255.255	用户可用的ASM临时组地址，全网范围内有效。
232.0.0.0~232.255.255.255	用户可用的SSM临时组地址，全网范围内有效。
239.0.0.0~239.255.255.255	用户可用的ASM临时组地址，仅在特定的本地管理域内有效，称为本地管理组播地址。

 Page14 

224.0.0.0 到 224.0.0.255 为 IANA 预留的永久组地址，地址 224.0.0.0 保留不做分配，其它地址供路由协议进行拓扑查找和维护协议使用。该范围内的地址属于局部范畴，不论生存时间字段（TTL）值是多少，都不会被路由器转发；

224.0.1.0到231.255.255.255，233.0.0.0到238.255.255.255为用户可用的ASM临时组地址，在全网范围内有效；

232.0.0.0到232.255.255.255，为用户可用的SSM临时组地址，全网范围内有效。

239.0.0.0到239.255.255.255，用户可用的ASM临时组地址，仅在特定的本地管理域内有效，称为本地管理组播地址。本地管理组播地址属于私有地址，在不同的管理域内使用相同的本地管理组播地址不会导致冲突。

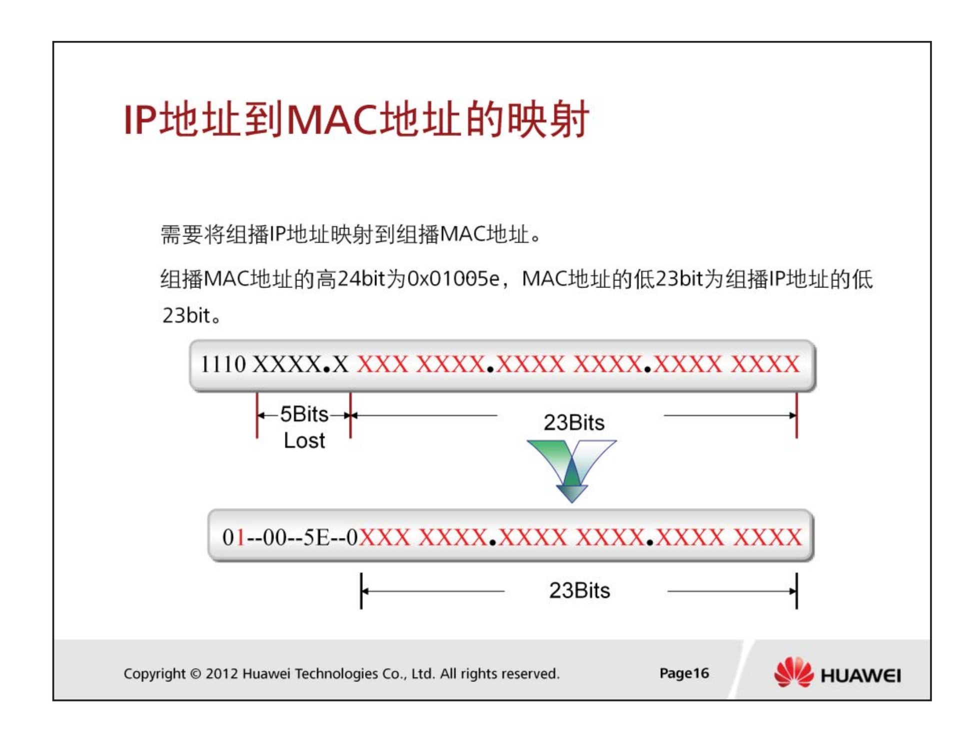




以太网传输单播IP报文的时候，目的MAC地址使用的是接收者的MAC地址。但是在传输组播报文时，传输目的不再是一个具体的接收者，而是一个成员不确定的组，所以使用的是组播MAC地址。

组播MAC地址用于在链路层上标识属于同一组播组的接收者。

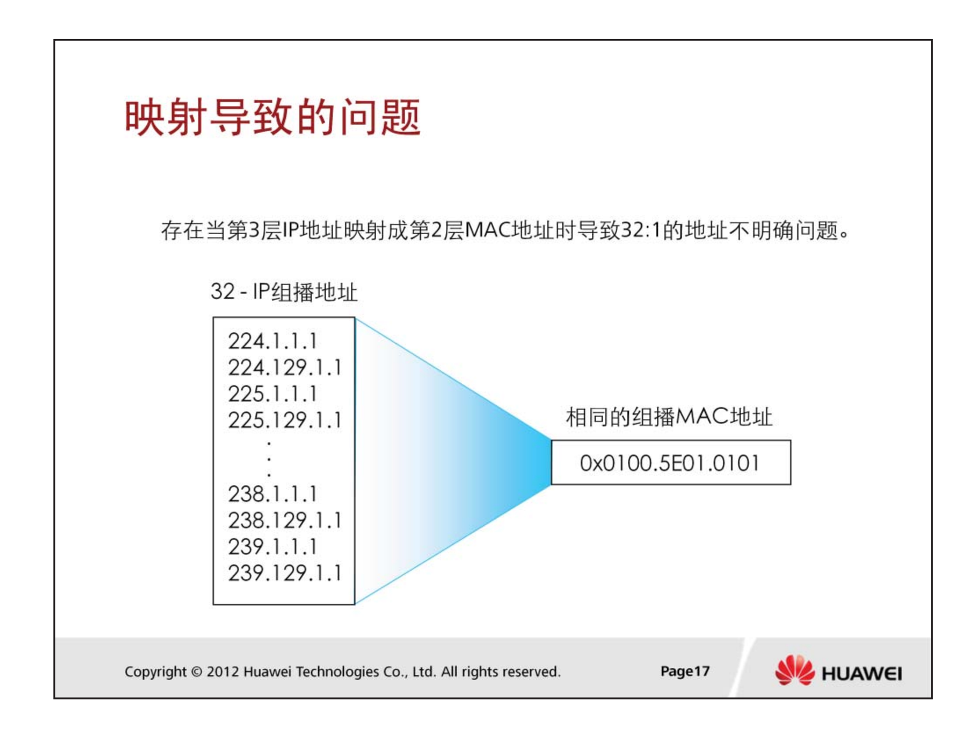
IANA规定，组播MAC地址的高24bit为0x01005e，第25bit固定为0，低23bit为组播IP地址的低23bit。



IP地址到MAC地址的映射如图所示。

组播MAC地址的高24bit为0x01005e。

IP组播地址的前4bit是1110，代表组播标识，而后28bit中有23bit被映射到MAC地址。



由于IP组播地址的前4bit是1110，代表组播标识，而后28bit中只有23bit被映射到MAC地址，这样IP地址中就有5bit信息丢失，直接的结果是出现了32个IP组播地址映射到同一MAC地址上。

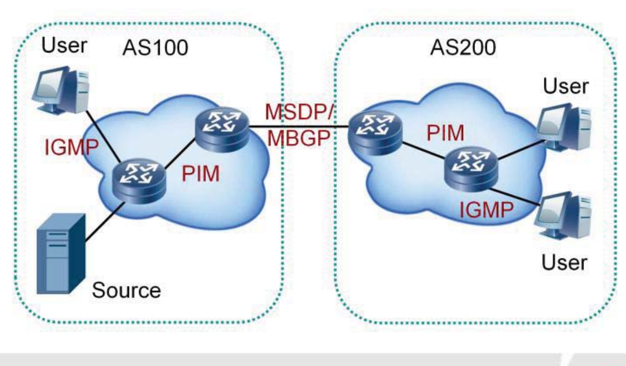


Page18


本章介绍组播相关的协议。组播组管理协议、组播路由协议和组播报文转发时两种不同的组播分发树，源路径树和共享树的概念。并分析两种分发树的区别。

组播相关协议

组播协议包括用于主机注册的组播组管理协议，和用于组播选路转发的组播路由协议。





组播协议包括用于主机注册的组播组管理协议，和用于组播选路转发的组播路由协议。各种组播协议在网络中的运用如图所示。

IGMP（Internet Group Management Protocol）在接收者主机和组播路由器之间运行，该协议定义了主机与路由器之间建立和维护组播成员关系的机制。

组播路由器之间运行组播路由协议，组播路由协议用于建立和维护组播路由，并正确、高效地转发组播数据包。

对于ASM模型，可以将组播路由分为域内和域间两大类。

域内组播路由协议用来在自治系统AS（Autonomous System）内发现组播源并构建组播分发树，将信息传递到接收者。域内组播路由协议包括：DVRMP、MOSPF、PIM。

DVRMP是距离矢量组播路由协议（Distance Vector Multicast Routing Protocol）是一种密集模式协议。该协议有跳数限制，最大跳数32跳。

MOSPF是OSPF路由协议的扩展协议。它通过定义新的LSA来支持组播。

PIM（Protocol Independent Multicast）是典型的域内组播路由协议，分为DM（Dense Mode）和SM（Sparse Mode）两种模型。当接收者在网络中的分布较为密集时，适用DM；较为稀疏时，适用SM。PIM必须和单播路由协议协同工作。

域间组播路由协议用来实现组播信息在AS之间的传递。

MSDP（Multicast Source Discovery Protocol）能够跨越AS传播组播源信息。

MPBGP （MultiProtocol Border Gateway Protocol）的组播扩展MBGP （Multicast BGP）能够跨越AS传播组播路由。

对于SSM模型，没有域内和域间的划分。由于接收者预先知道组播源的具体位置，因此可以借助PIM SM的部分功能直接创建组播传输路径。

本课程重点介绍域内组播路由协议。

组播分发树

什么是组播分发树?

- 用来描述IP组播报文在网络中经过的路径。

组播分发树的两个基本类型:

- 源路径树
  - 以组播源作为树根，将组播源到每一个接收者的最短路径结合起来构成的转发树。
- 共享树
  - 使用放在网络的某些节点的单独的公用根。根据组播路由协议，这个根常被称为汇合点（RP）或核心，因此，共享树也可以称为RPT。

 Page21



组播路由形成了一个从数据源到多个接收端的单向无环数据传输路径，即组播分发树。

组播分发树的两个基本类型：源路径树和共享树。

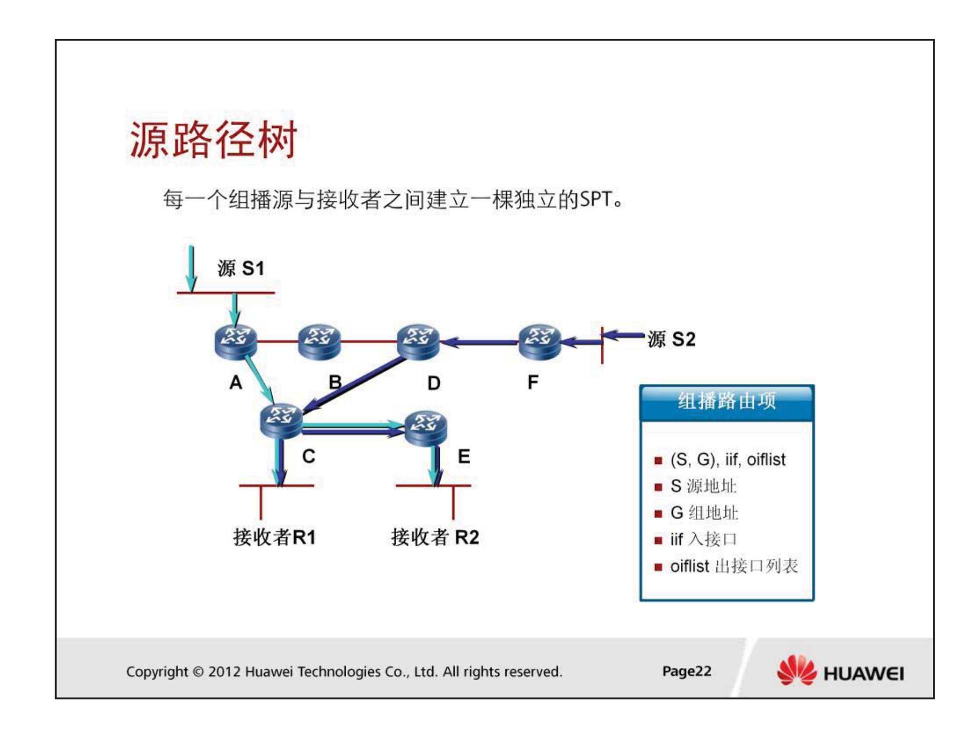
源路径树

以组播源作为树根，将组播源到每一个接收者的最短路径结合起来构成的转发树。由于源路径树使用的是从组播源到接收者的最短路径，因此也称为最短路径树（Shortest Path Tree，SPT）。对于某个组，网络要为任何一个向该组发送报文的组播源建立一棵树。

共享树

以某个路由器作为路由树的树根，该路由器称为汇集点（Rendezvous Point，RP），将 RP 到所有接收者的最短路径结合起来构成转发树。

使用共享树时，对应某个组，网络中只有一棵树。所有的组播源和接收者都使用这棵树来收发报文，组播源先向树根发送数据报文，之后报文又向下转发到达所有的接收者。



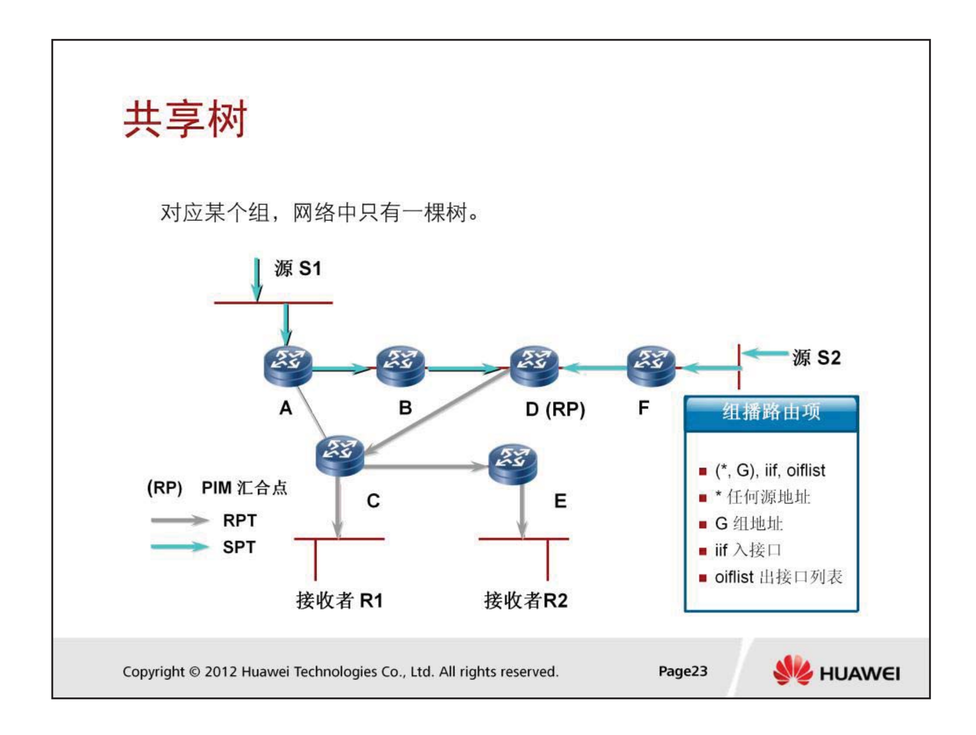
源路径树是以组播源作为树根，将组播源到每一个接收者的最短路径结合起来构成的转发树。

源路径树使用的是从组播源到接收者的最短路径，也称为最短路径树（shortest path tree，SPT）。对于某个组，网络要为任何一个向该组发送报文的组播源建立一棵树。

本例中有两个组播源（源S1和源S2），接收者R1和R2。所以本例中有两棵源路径树，分别是

S1—A---C（R1）-----E（R2）

S2---F----D---C（R1）------E（R2）



共享树以某个路由器作为路由树的树根，该路由器称为汇集点（Rendezvous Point，RP），将 RP 到所有接收者的最短路径结合起来构成转发树。使用共享树时，对应某个组，网络中只有一棵树。所有的组播源和接收者都使用这棵树来收发报文，组播源先向树根发送数据报文，之后报文又向下转发到达所有的接收者。

本例中两个源S1和S2共享一颗树 D（RP）----C（R1）----E（R2）

不同分发树的比较

源路径树（SPT）

- 路径最优，延迟最小，占用内存较多

共享树（RPT）

- 路径不是最优的，引入额外的延迟，占用内存较少

 Page24 

源路径树的优点是能构造组播源和接收者之间的最短路径，使端到端的延迟达到最小；但是付出的代价是，在路由器中必须为每个组播源保存路由信息，这样会占用大量的系统资源，路由表的规模也比较大。

共享树的最大优点是路由器中保留的状态可以很少，缺点是组播源发出的报文要先经过 RP，再到达接收者，经由的路径通常并非最短，而且对 RP 的可靠性和处理能力要求很高。



本章介绍组播转发RPF机制的原理。

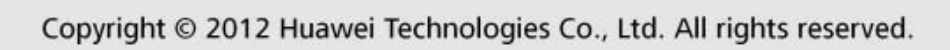


组播数据转发

组播路由和单播路由是相反的

- 单播路由关心数据报文要到哪里去。
- 组播路由关心数据报文从哪里来。

组播路由使用“反向路径转发”机制(RPF, Reverse Path Forwarding)



Page26


单播报文的转发过程中，路由器并不关心源地址，只关心报文中的目的地址，通过目的地址决定向哪个接口转发。

在组播中，报文是发送给一组接收者的，这些接收者用一个逻辑地址标识。路由器在接收到报文后，必须根据源和目的地址确定出上游（指向组播源）和下游方向，把报文沿着远离组播源的方向进行转发。这个过程称作 RPF（Reverse Path Forwarding，逆向路径转发）。

反向路径转发RPF

什么是RPF?

- 路由器收到组播数据报文后，只有确认这个数据报文是从自身连接到组播源的接口上收到的，才进行转发，否则丢弃。

RPF检查

- 在单播路由表中查找到组播报文源地址的路由
  - 如果该路由的出接口就是组播报文的入接口，RPF检查成功
  - 否则RPF检查失败，报文丢弃。

 Page27



RPF 执行过程中会用到原有的单播路由表以确定上游和下游的邻接结点。只有当报文是从上游邻接结点对应的接口（称作 RPF 接口）到达时，才向下游转发。

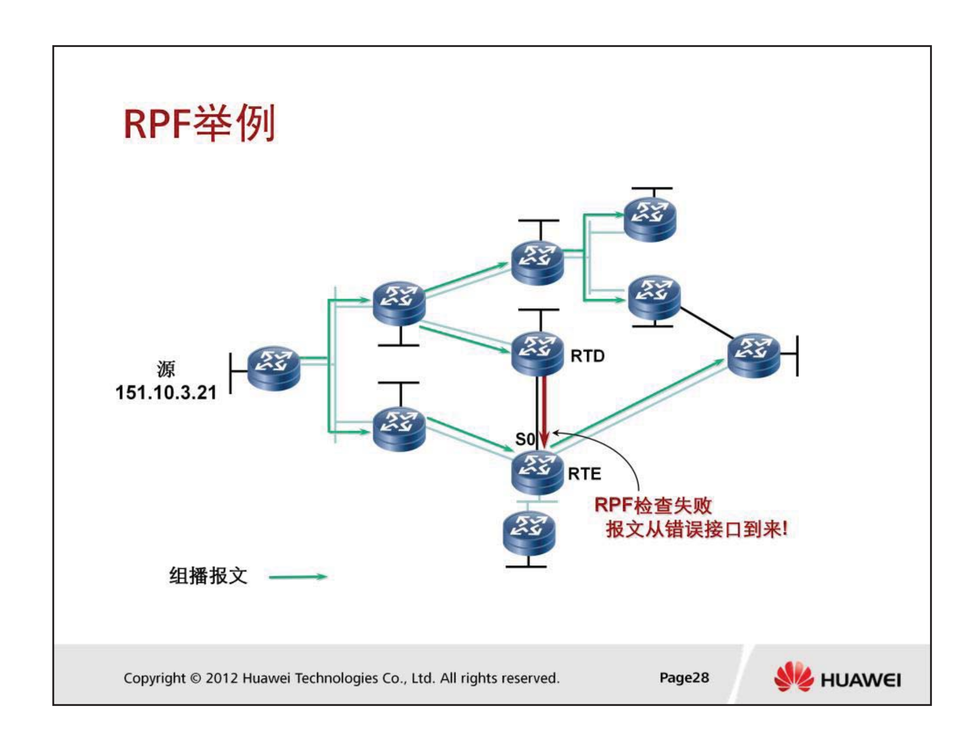
RPF 的作用除了可以正确地按照组播路由的配置转发报文外，还能避免由于各种原因造成的环路，环路避免在组播路由中是一个非常重要的问题。

RPF 的主体是 RPF 检查，路由器收到组播报文后，先对报文进行 RPF 检查，只有检查通过才转发，否则丢弃。RPF 检查过程如下:

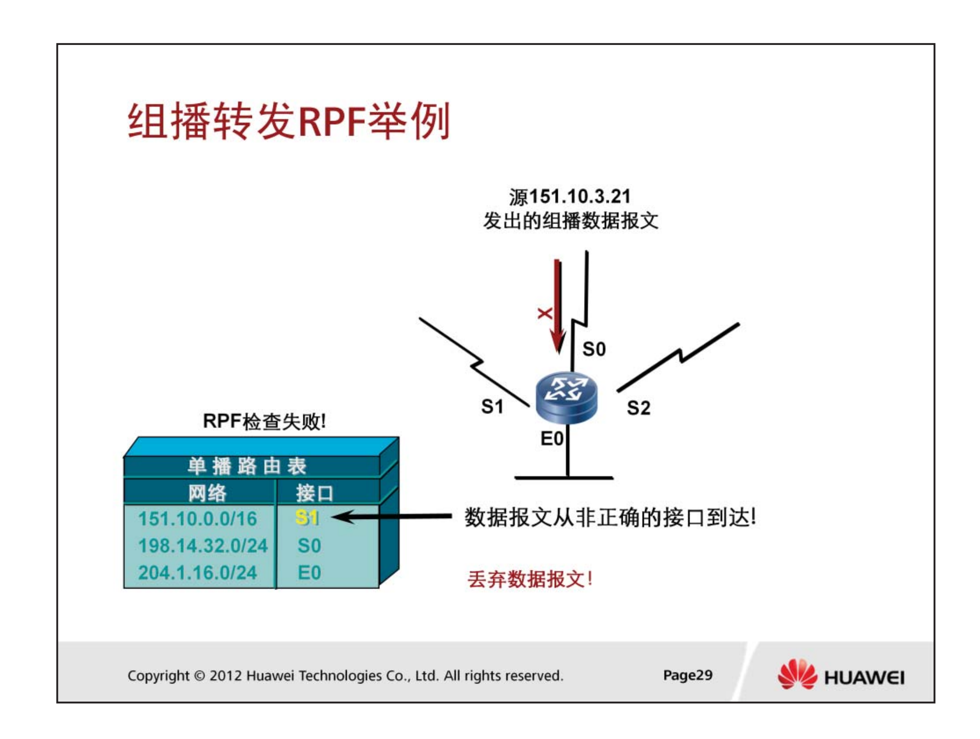
1）路由器在单播路由表中查找组播源或 RP 对应的 RPF 接口（当使用信源树时，查找组播源对应的 RPF 接口，使用共享树时查找 RP 对应的 RPF 接口），某个地址对应的 RPF 接口是指从路由器向该地址发送报文时的出接口;

2）如果组播报文是从 RPF 接口接收下来的，则 RPF 检查通过，报文向下游接口转发;

3）否则，丢弃该报文。



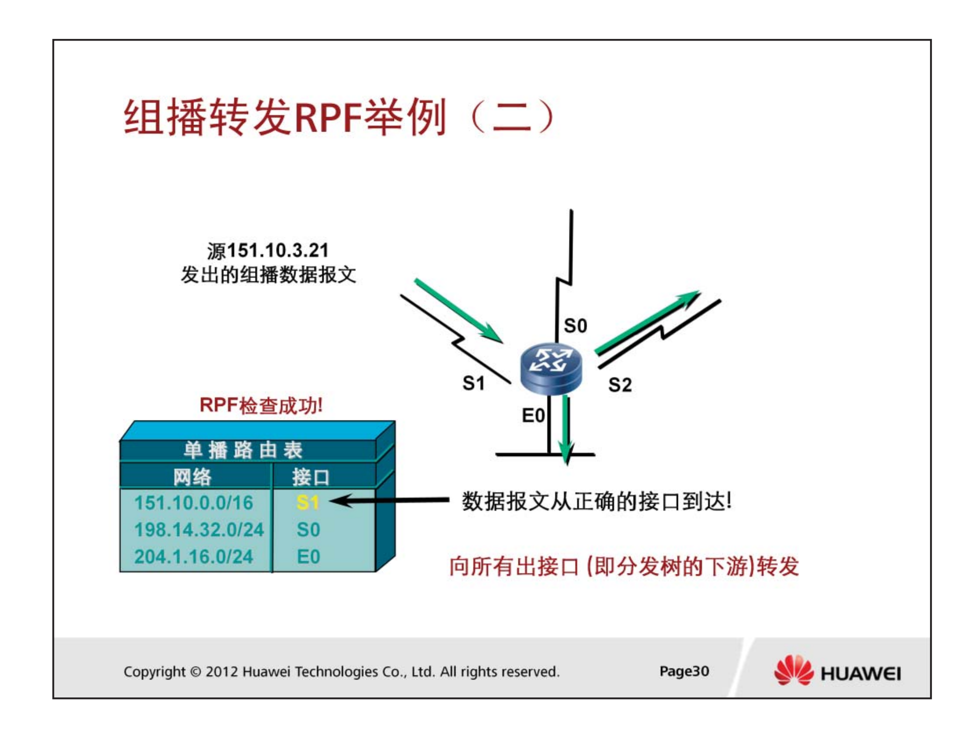
本例中路由器RTE从S0口接收到组播报文，于是该接口进行RPF检查。发现报文是从错误的接口收到的，于是RPF检查失败，RTE丢弃该报文。



RPF检查的过程实际上是查找单播路由表的过程。

路由器接收到组播报文后，后查找单播路由表，检查到达组播源的出接口是否与接收到组播报文的入接口相一致。如果一致则认为合法，如果不一致则认为从错误接口收到报文，RFP检查失败，丢弃报文。

如图所示的本例中，路由器的S0接口接收到组播报文。于是启动RPF检查。通过查找单播路由表，发现到达路由源151.10.0.0/16的出接口为S1，与接收口S0不一致，于是RPF检查失败，并认为是从错误接口收到组播报文，丢弃该报文。



与前一例相同的组网图，路由器从S1接口接收到组播报文，同样进行RPF检查。对比单播路由表，发现到达组播源的出接口是S1，与接收口一致，于是正常接收该报文，并向组播分树中的下游接口转发组播报文。



问 题

什么是组播?

组播地址结构?

组播相关协议?

什么是组播分发树? 组播分发树的类型?

组播转发机制RPF原理?

 Page31 

1、什么是组播?

组播组使用一个IP组播地址标识，接收者A和C两个信息接收者加入该组播组后从而可以接收发往该组播组的数据。组播源仅发送一次信息，借助组播路由协议为组播数据包建立树型路由，被传递的信息在尽可能远的分叉路口才开始复制和分发。相同的组播数据流在每一条链路上最多仅有一份。

2、组播地址结构?

D类组播地址范围是从224.0.0.0到239.255.255.255 。根据地址有效性分为：永久组地址和临时组地址。永久组地址是IANA为路由协议预留的组播地址，标识一组特定的网络设备。永久组地址保持不变，组成员的数量可以是任意的，甚至可以为零。临时组地址是为用户组播组临时分配的IP地址，组成员的数量一旦为零，即取消。

3、组播相关协议?

组播协议包括用于主机注册的组播组管理协议（IGMP）和用于组播选路转发的组播路由协议。组播路由协议分为域内组播路由协议和域间组播路由协议。域内组播路由协议有PIM-SM、PIM-DM、DVMRP，域间组播路由协议有MSDP、MBGP。

4、什么是组播分发树? 组播分发树的类型?

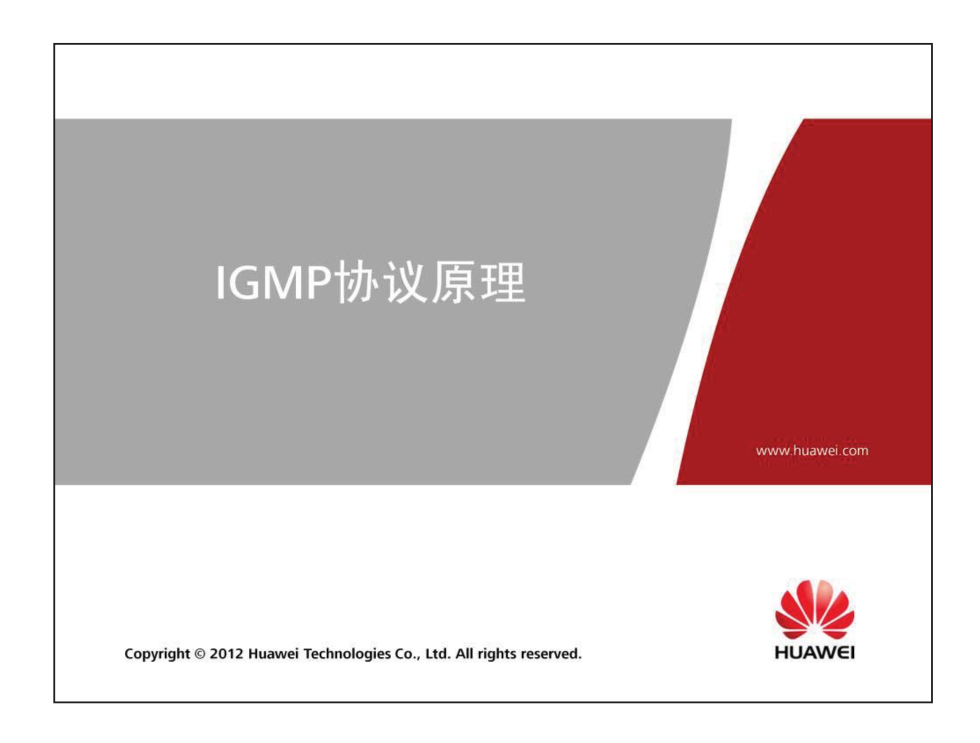
组播路由形成了一个从数据源到多个接收端的单向无环数据传输路径，

即组播分发树。组播分发树有两类：源路径树以组播源作为树根，将组播源到每一个分发树。组播分发树有两类：源路径树以组播源作为树根，将组播源到每一个接收者的最短路径结合起来构成的转发树，也称为最短路径树。共享树以某个路由器作为路由树的树根，RP 到所有接收者的最短路径结合起来构成转发树。

5、组播转发机制RPF原理?

RPF执行过程中会用到原有的单播路由表以确定上游和下游的邻接结点。只有当报文是从上游邻接结点对应的接口（称作 RPF 接口）到达时，才向下游转发。具体过程如下：路由器接收到组播报文后在单播路由表中查找到组播报文源地址的路由，如果该路由的出接口就是组播报文的入接口，RPF检查成功，否则RPF检查失败，报文丢弃。


谢谢
www.huawei.com


IGMP协议原理
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

IGMP（Internet Group Management Protocol）作为因特网组管理协议，是TCP/IP协议族中负责IP组播成员管理的协议，它用来在IP主机和与其直接相邻的组播路由器之间建立、维护组播组成员关系。

Page1



IGMP（Internet Group Management Protocol）作为因特网组管理协议，是TCP/IP协议族中负责IP组播成员管理的协议，它用来在IP主机和与其直接相邻的组播路由器之间建立、维护组播组成员关系。该协议在接收者主机和组播路由器之间运行，定义了主机与路由器之间建立和维护组播成员关系的机制。



培训目标

学完本课程后，您应该能：

- 理解IGMP协议原理
- 掌握IGMP配置
- 了解IGMP各版本区别
- 理解IGMP Snooping原理
- 掌握IGMP Snooping的基本配置

 Page2





本章介绍IGMP协议基本概念，IGMP发展情况。

IGMP协议介绍

IGMP 协议运行于主机和与主机直接相连的组播路由器之间。

IGMP工作机制：

- 接收者主机向所在的共享网络报告组成员关系。
- 查询器周期性地向该共享网段发送组成员查询消息。
- 接收者主机接收到查询消息后进行响应以报告组成员关系。
- 网段中的组播路由器依据接收到的响应来刷新组成员的存在信息。

Page4



IGMP协议是IP组播在末端网络上使用的主机对路由器的信令机制，分为两个功能部分：主机侧和路由器侧。

IGMP工作机制如下所述：

1.接收者主机向所在的共享网络报告组成员关系。

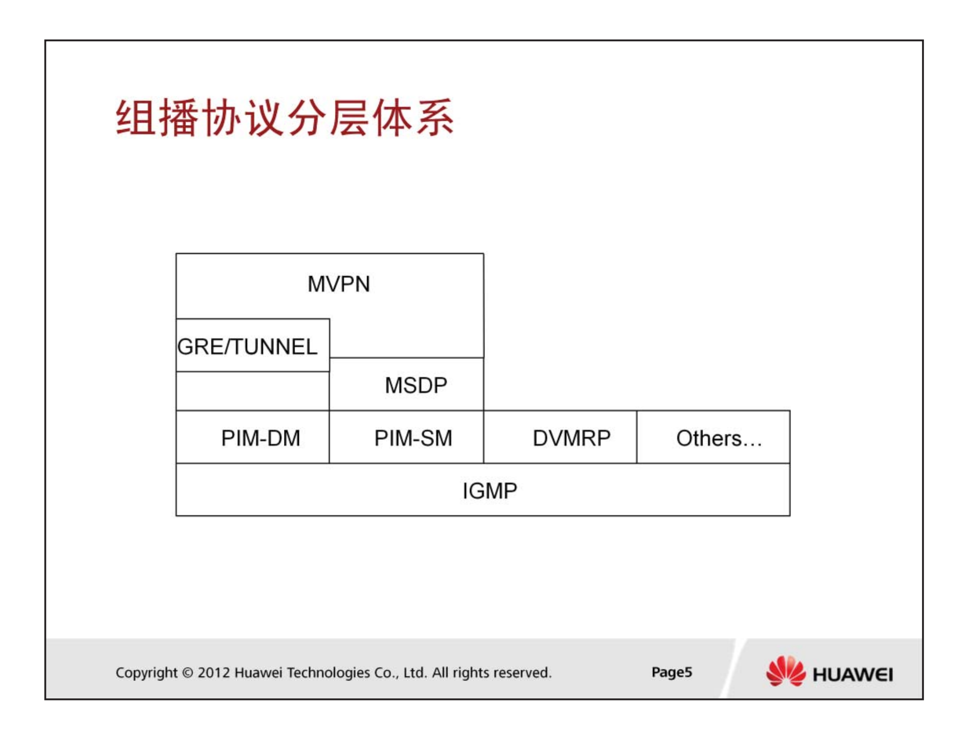
2.处于同一网段的所有使能了IGMP功能的组播路由器选举出一台作为查询器（查询器在不同的版本中有不同的选举机制），查询器周期性地向该共享网段发送组成员查询消息。

3.接收者主机接收到该查询消息后进行响应以报告组成员关系。

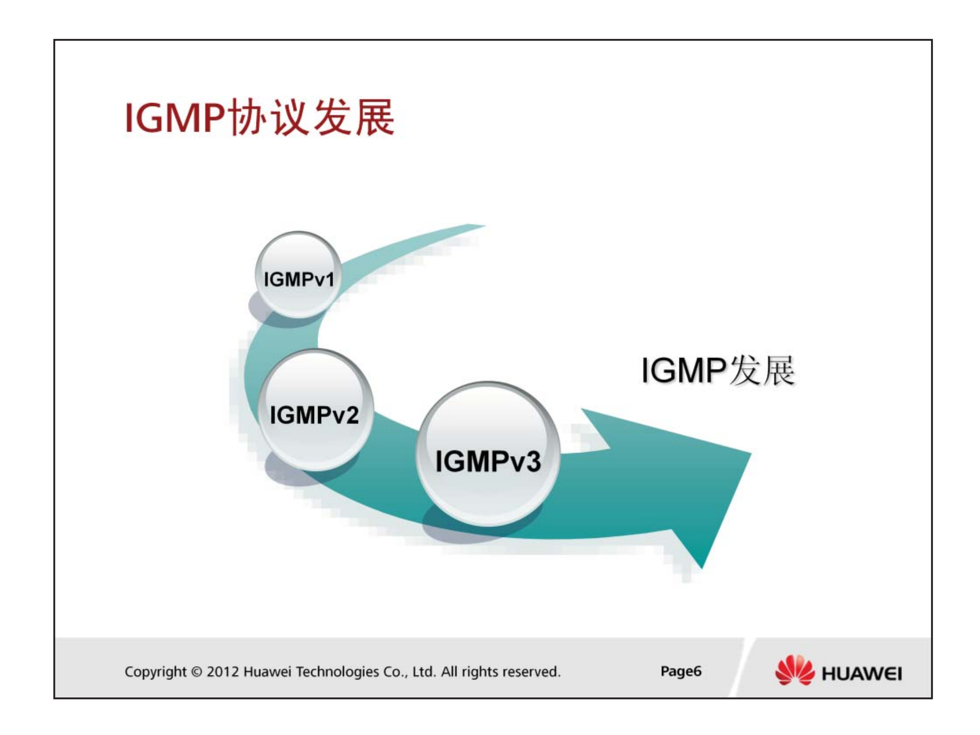
4.网段中的组播路由器依据接收到的响应来刷新组成员的存在信息。如果超时无响应，组播路由器就认为网段中没有该组播组的成员，从而取消相应的组播数据转发。

所有参与组播传输的接收者主机必须应用IGMP协议。主机可以在任意时间、任意位置、成员总数不受限制地加入或退出组播组。

支持组播的路由器不需要也不可能保存所有主机的成员关系，它只是通过IGMP协议了解每个接口连接的网段上是否存在某个组播组的接收者，即组成员。而各主机只需要保存自己加入了哪些组播组。



从此体系结构中可以知道IGMP处于组播协议的最底层，是整个组播协议体系的基础。在组播协议中，只有IGMP协议直接与主机联系，运行IGMP的路由器负责管理组成员主机的加入、离开，通过维护用户数据，发送组播数据到主机。



到目前为止，IGMP有三个版本：IGMPv1版本、IGMPv2版本和IGMPv3版本。所有IGMP版本都支持ASM（Any-Source Multicast）模型。IGMPv3可以直接应用于SSM（Source-Specific Multicast）模型，而IGMPv1和IGMPv2则需要SSM-Mapping技术的支持。

IGMPv1（RFC1112）定义了基本的组成员查询和报告过程。

IGMPv2（RFC2236）在 IGMPv1 的基础上添加了组成员快速离开的机制。

IGMPv3增加的主要功能是成员可以指定接收或指定不接收某些组播源的报文。



本章介绍IGMP协议不同版本中组播成员的加入、查询、离开工作机制。不同版本的路由器与主机之间IGMP的互操性。并通过对比了解IGMP不同版本的优缺点。

IGMPv1报文格式



版本
- 版本字段包含IGMP版本标识，因此设置为1。

类型
- 成员关系查询 （0x11）
- 成员关系报告 （0x12）

组地址
- 当一个成员关系报告正被发送时，组地址字段包含组播地址。
- 当用于成员关系查询时，本字段为0，并被主机忽略。

 Page8 

IGMPv1报文格式如图所示，各字段表示如下：

Version：IGMP版本标识，版本1为1。IGMPv2的报文中没有该字段。

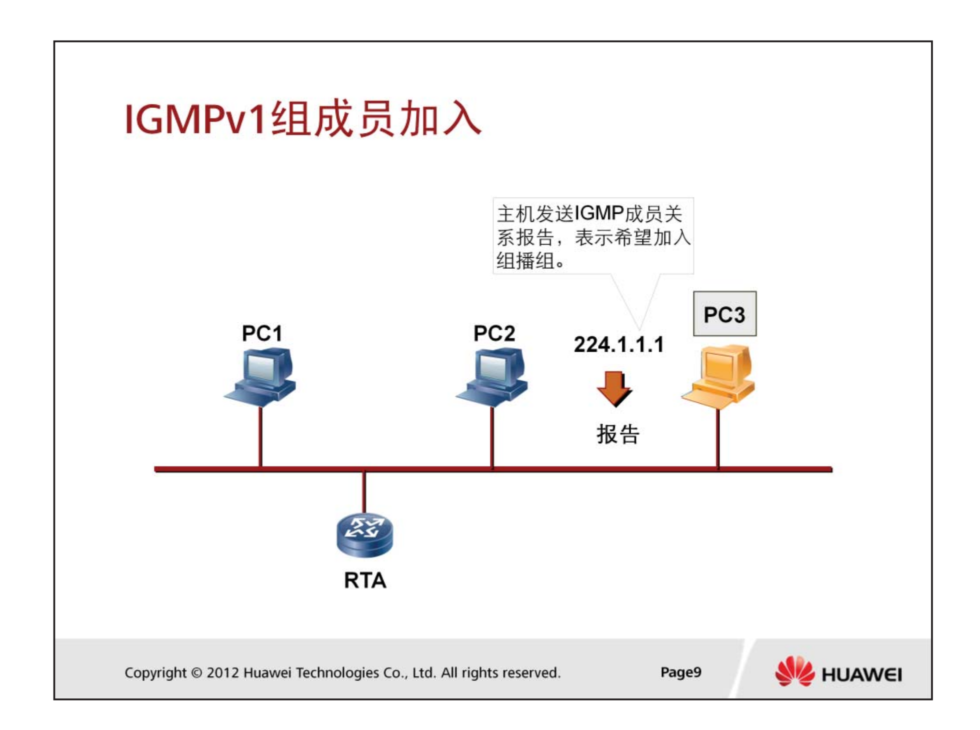
Type：类型字段。表示IGMP报文类型。

IGMPv1支持两种类型的报文：

- 成员关系查询（0x11）：路由器周期性的发送成员关系查询报文去查询是否有组播成员。默认查询周期为60秒。
- 成员关系报告（0x12）：成员关系报告用于表示主机想加入某个组播组。
- 成员关系报告的发送可以被动发送也可主动发送。
- 被动发送是指当主机收到成员关系查询消息后如果对某个组播组感兴趣想加入组播组时发送成员关系报告。
- 主动发送是指如果主机想加入某个组播组时，可以不用等待成员关系查询报文，而主动地发送成员关系报告。

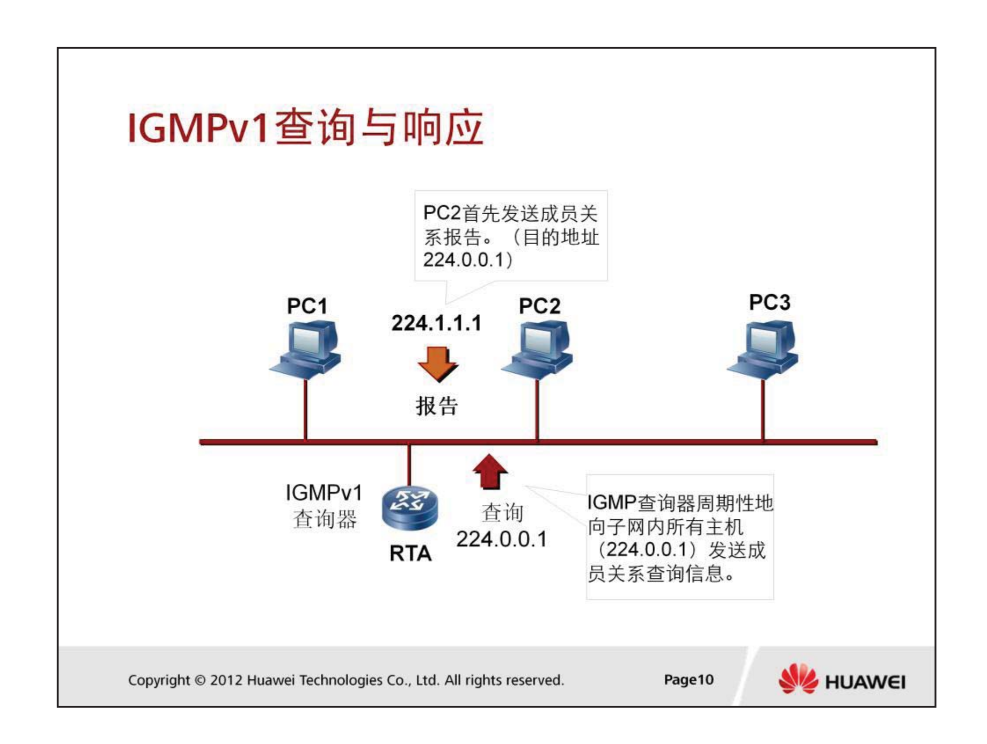
组地址：不同类型的IGMP报文中组地址不同。

- 在成员关系报告报文中，组地址为某个特定的组播地址。
- 在成员关系查询报文中，组地址为0。



主机主动发送IGMP成员关系报告报文，表示想加入一个组播组中。

该报文中组地址为主机想加入的组播组的地址。



IGMPv1的查询与响应过程如下：

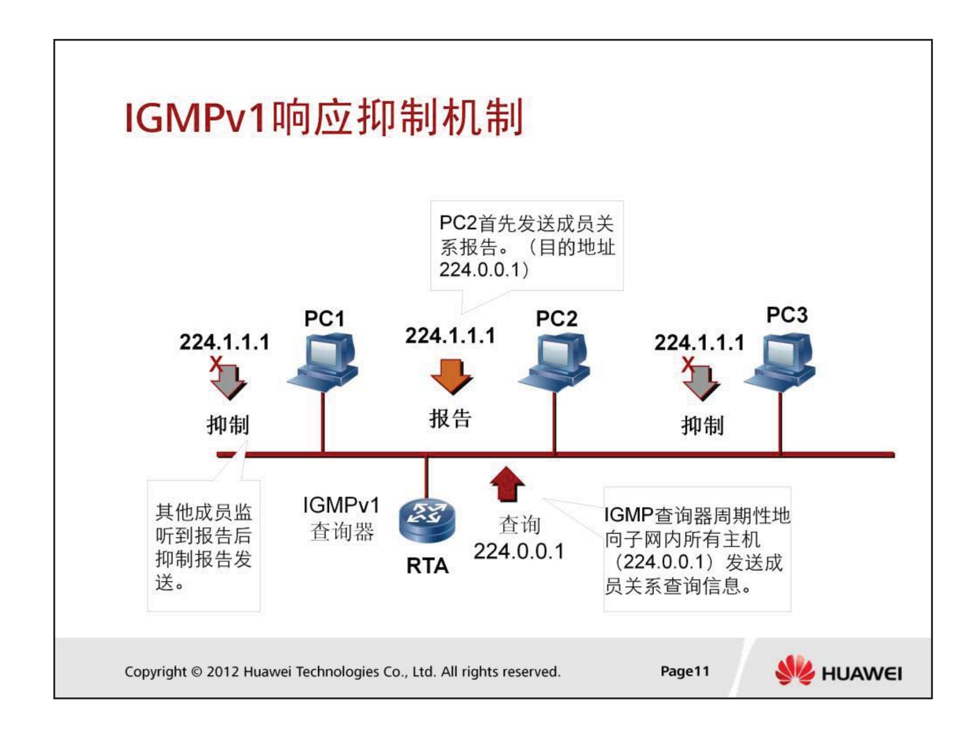
1、IGMP查询器周期性地向共享网段内所有主机以组播方式（目的地址为224.0.0.1）发送成员关系查询消息（组地址为0）。

2、网络内所有主机都接收到该查询消息，如果某主机（如PC1、PC2和PC3）对任意组播组G感兴趣，则以组播方式发送“成员关系报告”报文（其中携带组播组G的地址）来宣告自己将加入该组播组，假设PC2首先发送此报告。

3、经过查询/响应过程后，IGMP路由器了解到本网络内存在组播组G对应的接收者，生成（*，G）组播项并依此作为组播信息的转发依据。

*表示任意组播源，G表示某个组播组。

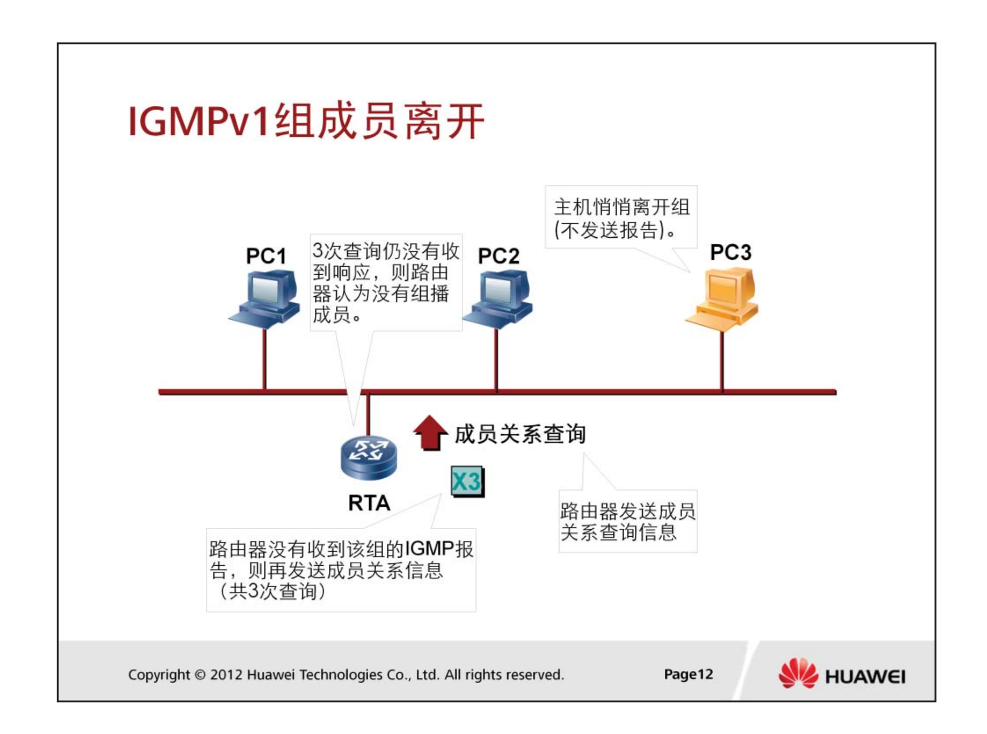
IGMPv1规定，当共享网络中有多台路由器时，由组播路由协议选举查询器。不同的组播路由协议有不同的选举机制。



IGMP成员关系查询报文是目的地址是224.0.0.1，就是说网段内所有的设备都会接收到该查询报文。但并不是所有接收到该报文的主机都会响应查询请求的。本例中只有一个主机会以成员关系报告报文响应，而其他主机则抑制成员关系报告的发送。

实际上主机收到IGMP成员关系查询时，会对它已经加入的每个组播组启动一个倒计数报告计时器。IGMPv1中计时器值固定使用10秒。计时器到时的主机则主动发送成员关系报告，组地址为该组播组地址，目的地址为224.0.0.1。于是网段内其它主机都会收到该成员关系报告报文，接收到成员关系报告报文的主机抑制成员关系报告的发送，并删除计时器。

当路由器周期性的发送成员关系查询报文时，每个主机都会再次启动计时器进行查询/响应/抑制。



由于IGMPv1版本没有定义专门离开组播组的消息，因此主机离开组时是默默离开不发送任何报文。而组播路由器如何知道用户已经离开组播组呢？IGMPv1主要是基于查询无响应进而超时的思路实现的。

成员悄悄离开组播组，不发送任何报文。路由器依旧周期性的发送成员关系查询报文，周期为60秒，当路由器发送3次成员关系查询报文都没有收到响应的成员关系报告报文时，路由器认为组内已经没有成员，不再向该网段转发组播报文。

IGMPv2报文格式



类型

- 成员关系查询（0x11）
  - 常规查询：用于确定哪些组播组是有效的，即该组是否还有成员在使用，常规查询地址由全零表示；
  - 指定组查询：用于查询特定的组播组是否还有组成员。
- 版本2成员关系报告（0x16）
- 版本1成员关系报告（0x12）
- 离开组消息（0x17）

Page13


IGMPv2报文与IGMPv1报文略有不同，取消了版本字段而相应增加了最大响应时间字段。

IGMPv2报文中有三种报文类型：

Type=0x11 成员关系查询报文，又分两种子类型：

- 常规查询：用于确定哪些组播组是有效的，即该组是否还有成员在使用，常规查询组地址全零；
- 特定查询：用于查询特定的组播组是否还有组成员。组地址为特定的组播地址。

Type=0x16 IGMPv2组成员关系报告。

- 为了和IGMPv1兼容，还有另外的一个附加的消息类别：
- 0x12 = IGMPv1成员报告。

Type=0x17 离开组消息，主机发送的离开报告。

IGMPv2报文格式

最大响应时间

- 以0.1秒为单位
- 默认值是100，即10秒。

校验和

组地址

- 在成员查询消息中，发送一个常规查询时组地址域设为0，当发送一个特定组查询时，则应设置组的地址。
- 在成员报告或离开组的消息中，组的地址域保留了要报告或要离开的地址。

Page14



最大响应时间字段，仅用于组成员关系查询。表示主机响应查询返回报告的时间范围。IGMPv1中没有该字段。

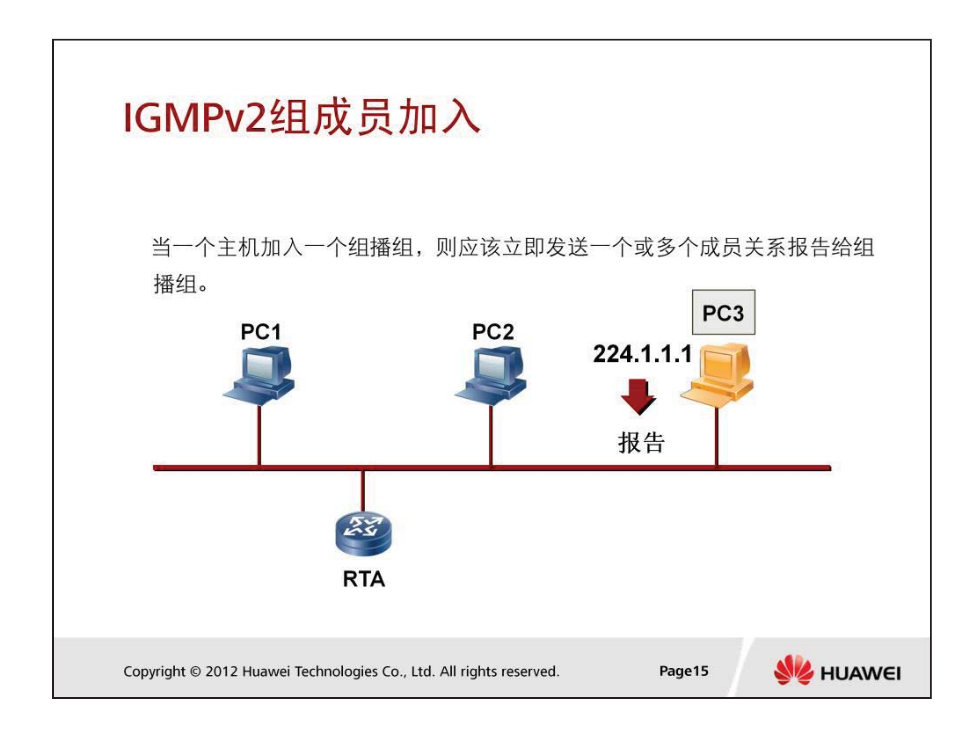
组地址字段：

发送常规查询时，组地址字段设置为零；

特定组查询时候，设置为要查询的组地址。

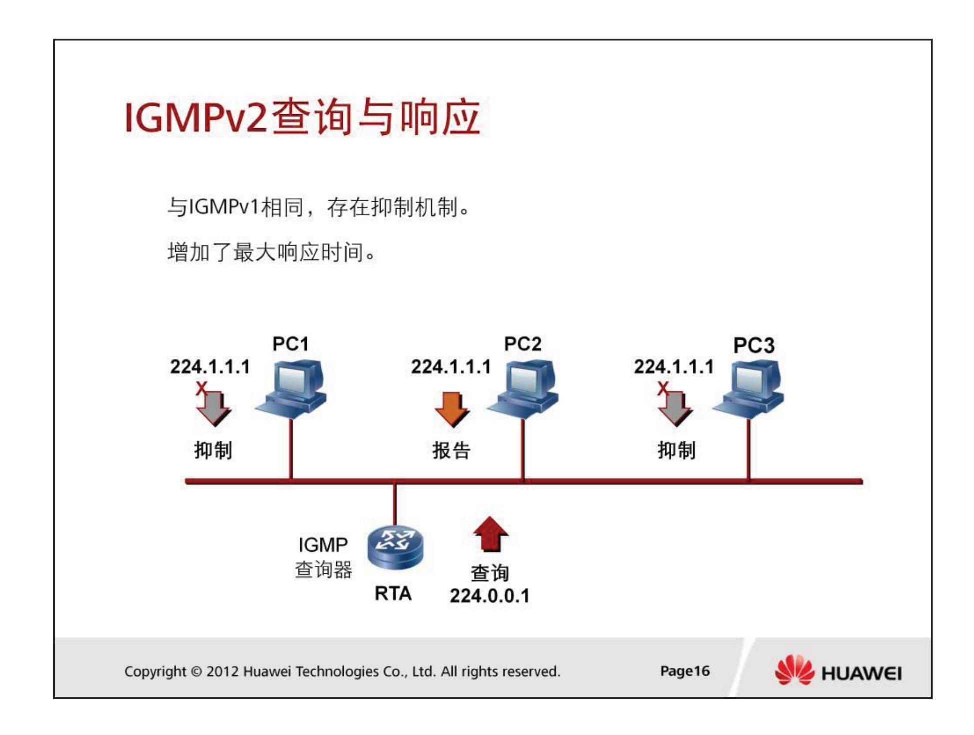
当主机成员发回组关系报告或是发送离开组消息时，本字段设置为目标组地址。

校验和是IGMP消息长度(IP包的整个有效负载)的16位检测，该域设为0。



当一个主机首次加入组播组时，主机立即发送成员关系报告报文。初始的成员报告可能会丢失或会受到损害，为了防止此种情况，推荐在短的间隔时间内报告一次或两次（RFC2326推荐的时间间隔为10秒)。

IGMPv2主机也支持IGMPv1的成员关系报告报文。

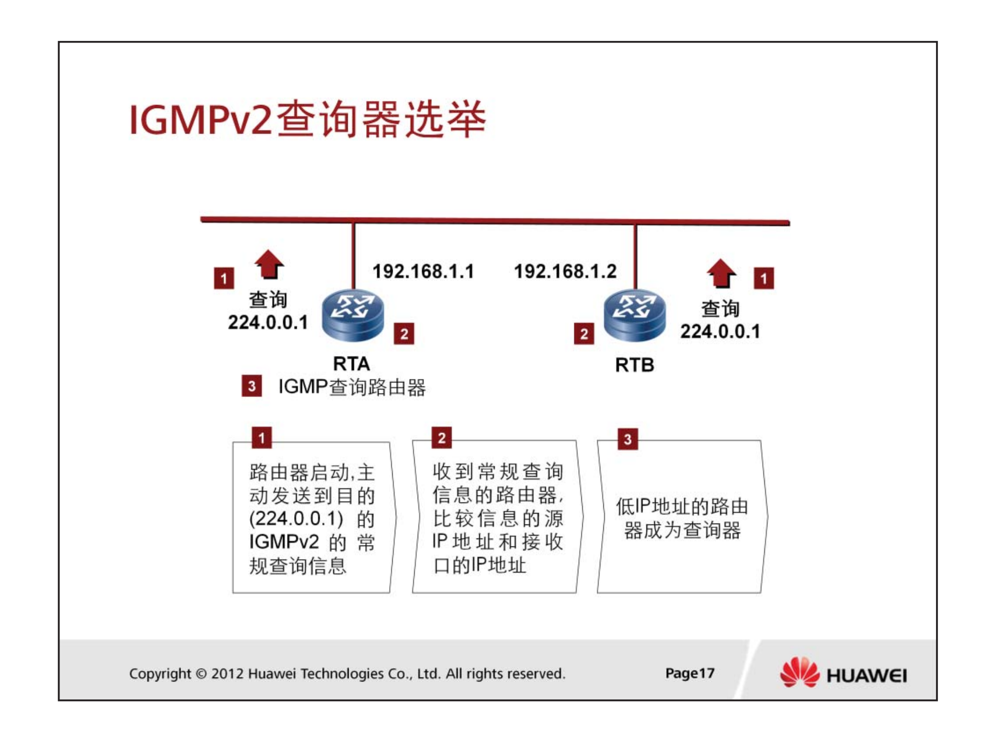


IGMPv2中增加了最大响应时间字段。前面介绍了主机收到成员关系查询报文时，会为每个已经加入的组播组启动一个计时器。计时器到期的主机才会发送IGMP成员关系报告报文来响应路由器的查询。在v2中该计时器的值为“1~最大响应时间”之间的一个随机值。

IGMPv2版本增加最大响应时间字段，以动态地调整主机对组查询报文的响应时间。

在IGMPv1版本中，组播路由器发起的查询是针对该网段下的所有组播组，这种查询被称为普遍组查询。

IGMPv2版本中，在普遍组查询之外增加了特定组的查询，这种查询报文的目的IP地址为某个组播组的IP地址，报文中的组地址字段也为该组播组的IP地址，网络中属于该组播组成员的主机才会进行响应，这样就避免了属于其它组播组成员的主机发送响应报文。



对于一个网段上有多个组播路由器的共享网段，此网段下运行IGMP的路由器都能从主机那里收到成员关系报告消息，但是只需要一个路由器发送成员资格查询消息，所以这就需要一个路由器选举机制来确定一个路由器作为查询器。

只有查询器才会发送成员关系查询报文。在IGMPv1版本中，查询器的选择由组播路由协议决定；

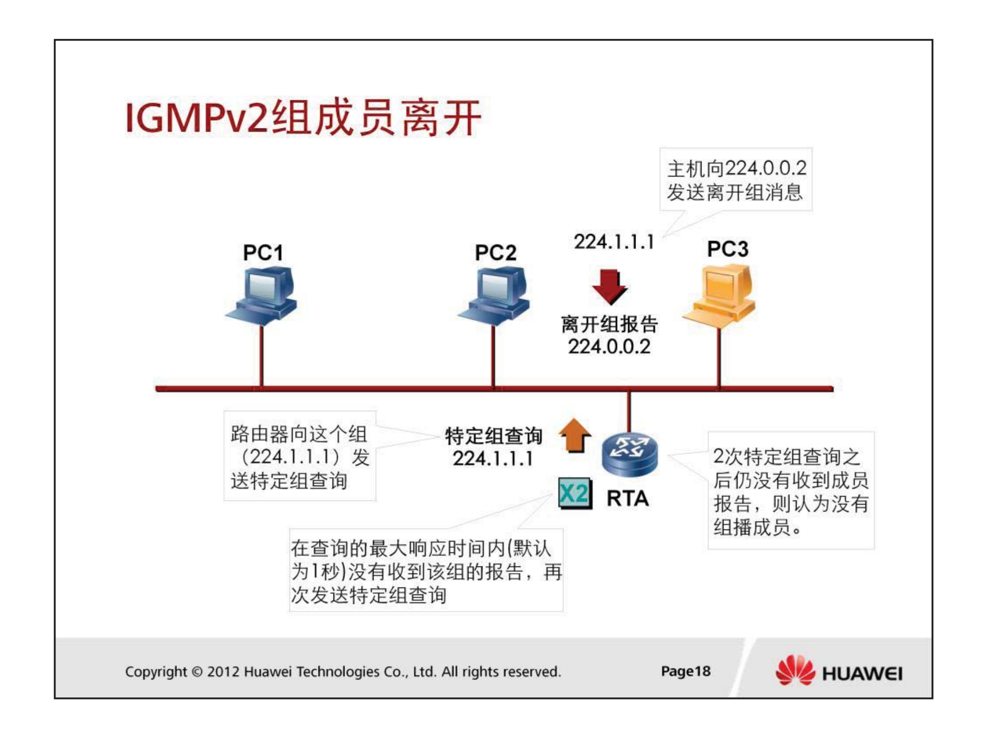
IGMPv2版本对此做了改进，规定同一网段上有多个组播路由器时，具有最小IP地址的组播路由器被选举出来充当查询器。

路由器启动，主动发出到目的地（224.0.0.1）的IGMPv2常规查询信息

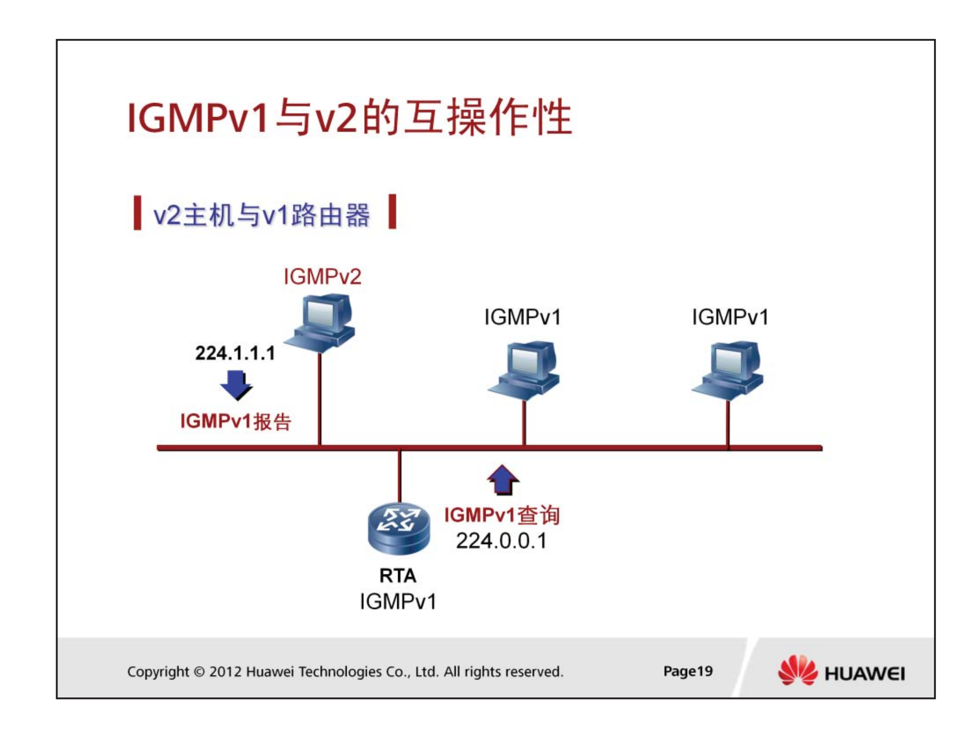
。

收到常规查询信息的路由器，会把此信息的源IP地址和接收口的IP地址作比较，拥有最低IP地址的路由器被选举为IGMP查询路由器。

查询器也会有失效的时候，当查询器失效时，另一路由器成为查询器。所以非查询路由器会启动一个查询计时器，周期检查IGMP查询路由器的状态，缺省情况下120秒。该值可以通过命令 timer other-querier-present interval 修改。



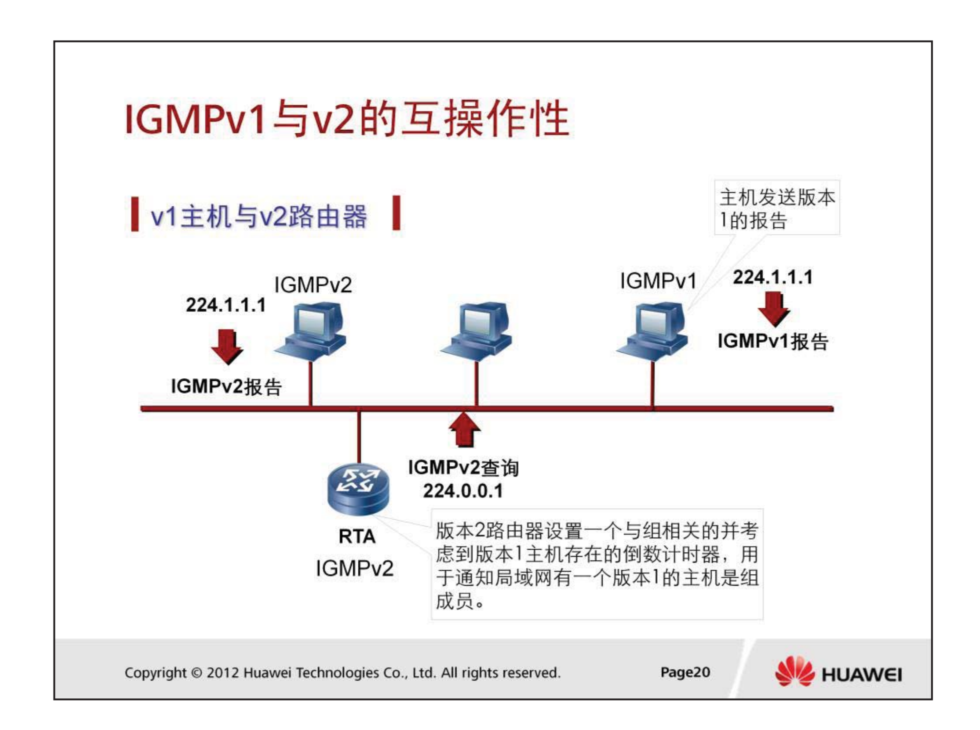
在IGMPv1版本中，主机悄然离开组播组，不会向任何组播路由器发出任何通知。造成组播路由器只能依靠响应超时来获知组播成员的离开。而在v2版本中，当一个主机决定离开一个组播组时，它会向网络中所有组播路由器以组播方式（224.0.0.2）发送离开组的消息，为了明确该组播组中是否还包含其它成员主机，该组播路由器会向网络中发送特定组查询消息。在查询的最大响应时间内（默认为1秒）没有收到该组的报告，则再次发送特定组查询。2次特定组查询后仍没有收到成员报告，则认为组播成员全部离开。



版本1路由器把IGMPv2报告看作无效的IGMP信息类型并且忽略它。当版本1路由器作为有效的IGMP查询器的时候，版本2的主机必须发送IGMPv1报告。

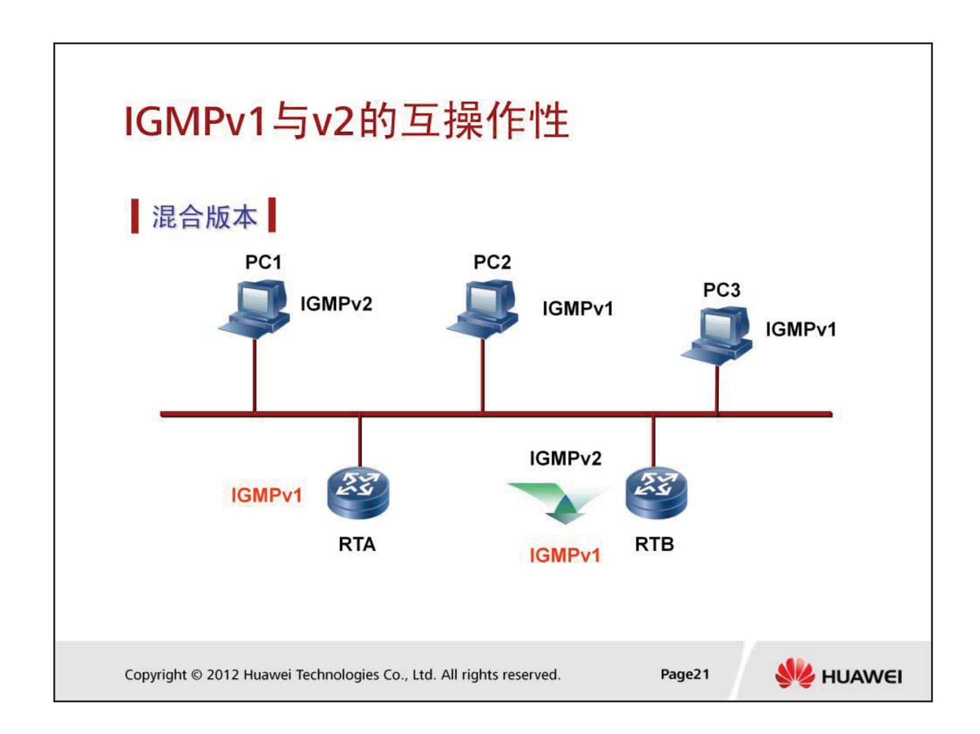
当版本2的主机检测出IGMP查询器是版本1的路由器时，它必须始终用IGMPv1报告做出响应。在这种情况下，版本2的主机也可以抑制发送离开组信息。为了维护本接口的状态，无论何时IGMPv1查询在接口处被收到，版本2主机会启动一个400秒的倒数计时器，当另一个IGMPv1查询被收到时，计时器被复位。如果计时器到时，此接口恢复成为IGMPv2接口并且IGMPv2信息被再次发送。

版本2主机必须允许它的成员关系报告被IGMPv1或IGMPv2抑制。

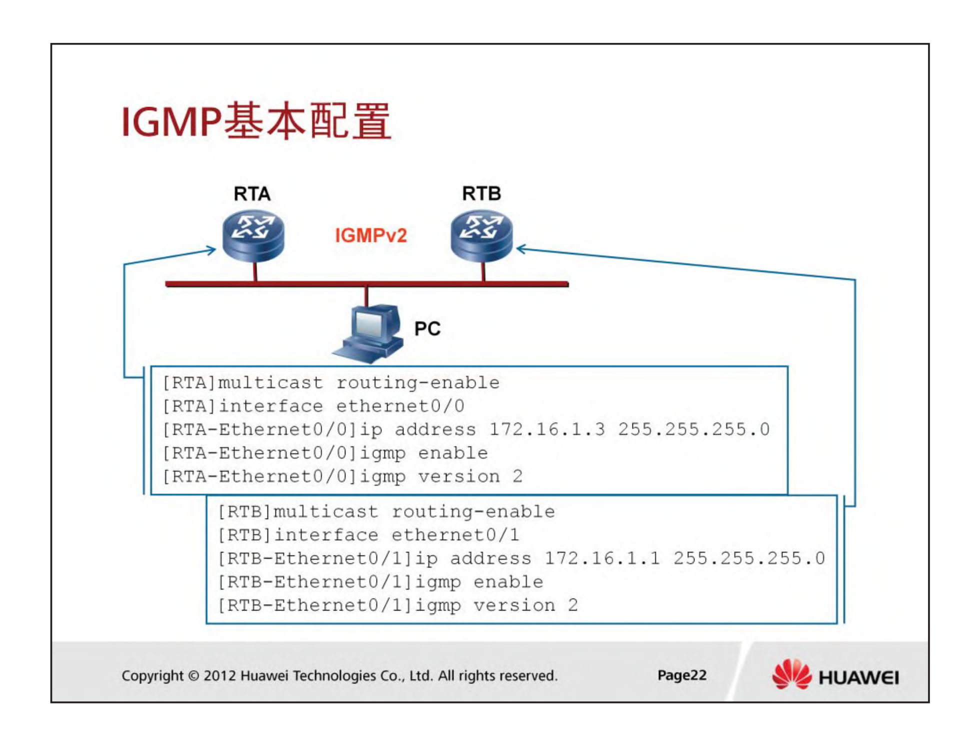


如果某个版本2的路由器是IGMP查询器，同时在局域网中版本1的主机也是同一组成员，那么该组的IGMPv1报告总是被收到，因为IGMPv2报告在版本1的主机中不会被抑制，版本1的主机不兼容版本2的报告，因此忽略它们。

无论何时，只要版本1的主机成为组成员，IGMPv2离开过程都将被搁置。



如果一些运行IGMP版本1的路由器存在于子网中，那么必须强制性地为子网中的所有路由器配置IGMPv1以便正常使用。



IGMP应用在路由器与用户相连的网段，在路由器和用户主机上都需要运行IGMP。这部分只介绍如何在路由器上配置IGMP。

配置IGMP之前，必须先使能IP组播路由。IP组播路由是配置一切组播功能的前提。如果停止IP组播路由，组播所有相关配置将被删除。

[Huawei]multicast routing-enable 使能IP组播路由。

在连接用户主机的接口上使能IGMP，由于不同版本的IGMP协议报文不相同，因此需要为路由器和主机配置匹配的版本。

[Huawei-Ethernet0/0]igmp enable 在接口下使能IGMP。

为了让接口所连接网络上的主机加入指定的组播组，并接收这些组的报文，可以在对应接口上设置一个ACL规则作为过滤器，以限制接口所服务的组播组范围。

IGMP的版本配置可以在两种情况下进行:

1、接口:

[Huawei-Ethernet0/0]igmp version 2

针对接口的配置优于全局配置，接口未配置时则继承全局配置。

2、全局:

[Huawei]igmp 全局使能IGMP。

[Huawei-igmp]version 2 配置全局IGMP的版本。

IGMP配置验证

```
<RTD>display igmp interface
Ethernet0/0 (172.16.1.1):
IGMP is enabled
Current IGMP version is 2
Value of query interval for IGMP(in seconds): 60
Value of other querier time out for IGMP(in seconds): 120
Value of maximum query response time for IGMP(in secot for
IGMP: 2
Value of startup query interval for IGMP(in seconds): 15
Value of last member query interval for IGMP(in seconds): 1
Value of query timeout for IGMP version 1(in seconds): 400
```

 Page23



display igmp interface 查看接口上IGMP配置和运行信息，如果在全局模式下使用该命令，则可以显示所有使能了IGMP协议的接口运行信息。如果在某一接口下使用则显示本接口的IGMP运行信息。

如果IGMP协议配置正确，则通过该命令可查看到使能接口的IP地址，IGMP协议版本，以及相应的IGMP的配置参数值。

Value of query interval for IGMP(in seconds): 60表示IGMP的普通组成员查询报告的时间间隔，默认值为60秒。

Value of other querier time out for IGMP(in seconds): 120表示IGMP查询器的超时时间是120秒。

IGMP协议涉及较多的配置参数，这里不一一列举。有兴趣的读者可以参考VRP操作手册。

IGMP配置验证

```
[RTB]display igmp group
 Total 2 IGMP groups reported on this router
Ethernet0/1 (172.16.1.1): Total 2 IGMP Groups reported:
  Group Address        Last Reporter   Uptime      Expires
  239.255.255.250      172.16.1.5      00:08:04    00:02:52
  224.1.1.1            00:03:00
```

 Page24 

display igmp group命令用来查看IGMP组播组信息，包括通过成员报告动态加入的组播组和通过命令行静态加入的组播组。

上图中粗体表示的是已经加入的组播组信息。

IGMPv3概述

IGMPv3在RFC 3376中说明（尚未得到广泛支持）。

服务于Source Specific Multicast（SSM）模型。

允许主机指定接收某些网络发送的某些组播组。

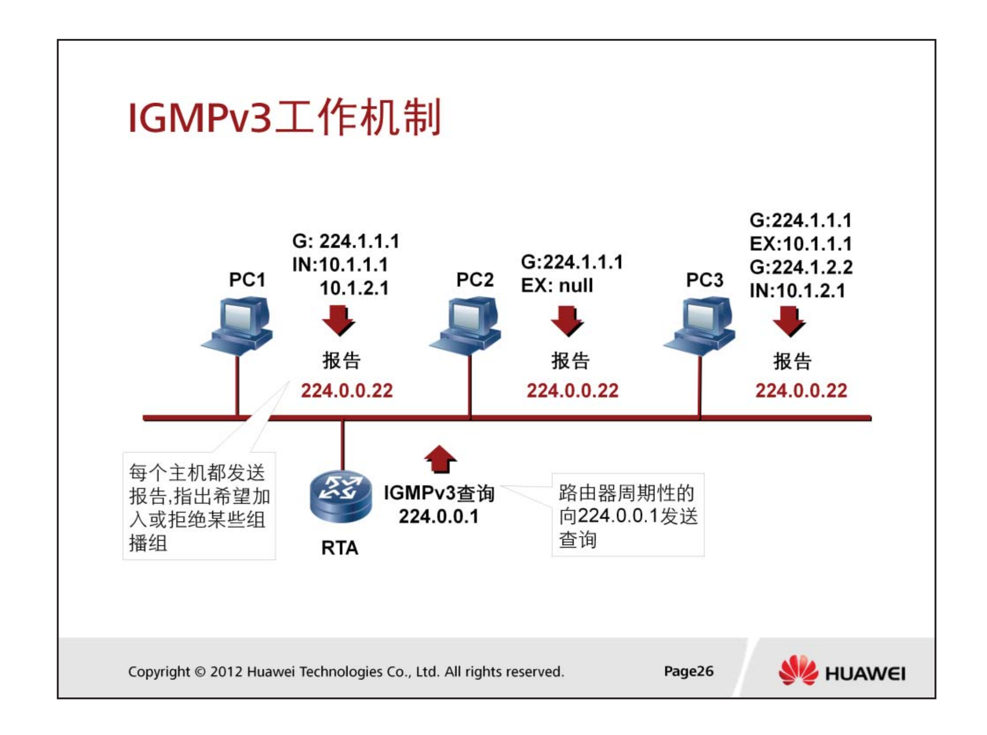
增加了主机的控制能力，不仅可以指定组播组，还能指定组播的源。

 Page25 

IGMPv3版本在兼容IGMPv1和v2版本基础上进一步增强了主机的控制能力，不仅可以指定加入的组播组G，还能明确要求从哪个指定组播源S接收信息，这也就是指定源组播功能。

如果主机仅需要获得某些特定源的信息，可以将IGMP报告中的Filter-Mode字段设置为Include模式，并在该报告中指定需要接收的组播源地址Sources，从而实现从指定源地址接收组播报文，鉴于描述方便可以表示为Include Sources（S1，S2，……）；

如果主机不想接收某些特定源的信息，则可以要求从除指定源外的所有其他源地址接收组播报文，在IGMP报告中标记为Exclude Sources（S1，S2，……）。

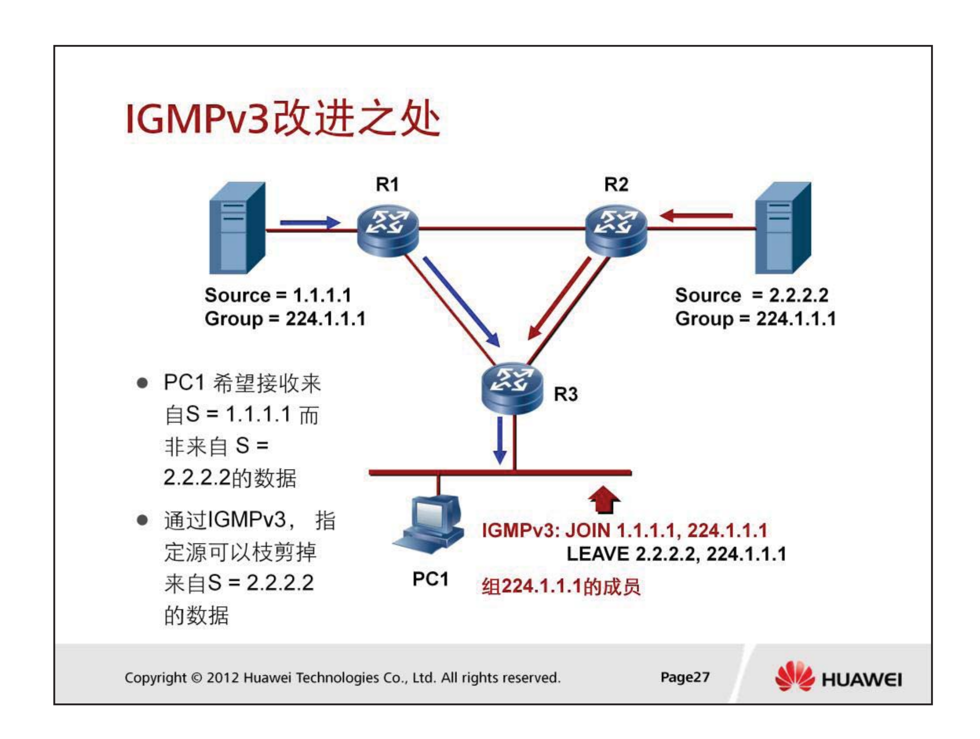


IGMPv1和v2版本的响应消息和查询消息具有相同的报文结构，即报文中仅包含组地址信息。IGMPv3响应消息包含的组地址为224.0.0.22，其中可以携带1个或多个组记录，在每个组记录中，包含组播组地址、数目不等的源地址信息。组记录可以分为多种类型，如：

当前状态记录：报告了接口的当前接收状态，分为Include和Exclude两种状态类型。Include表示包含指定源地址列表，Exclude表示包含除指定源地址列表外的所有源地址。

过滤模式改变记录：报告接口接收状态从Include状态切换到Exclude状态，或从Exclude状态切换到Include状态。

源地址列表改变记录：报告新源地址加入，或删除某源地址。



IGMPv3不仅支持IGMPv1版本的普遍组查询，支持IGMPv2版本的特定组查询，而且支持IGMPv3版本的指定源/组查询。在IGMP消息中携带组播源地址和多种控制字段（如查询器的强壮性系数、查询间隔等）。对于普遍组查询，既不携带组地址，也不携带源地址；对于特定组查询，携带组地址，但是不携带源地址；对于指定源/组查询，既携带组地址，而且还携带1个或多个源地址信息。

IGMP版本比较

	IGMPv1	IGMPv2	IGMPv3
查询器选举	依靠上层路由协议	自己选举	自己选举
成员离开方式	默默离开	主动发出离开报文	主动发出离开报文
指定组查询	不支持	支持	支持
指定源、组加入	不支持	不支持	支持

 Page28 HUAWEI

本页对比IGMP的三个不同版本。IGMPv1不支持查询器的选举，查询器的选举依靠组播路由协议来选举。



本章介绍IGMP Snooping协议原理和配置。

IGMP Snooping概述

IGMP Snooping解决组播报文在二层广播的问题。

IGMP Snooping运行在链路层，是二层以太网交换机上的组播约束机制，用于管理和控制组播组。

IGMP Snooping通过监听主机发出的IGMP报文，建立组播MAC地址表。

 Page30 

IGMP Snooping运行在数据链路层，用于管理和控制组播组，解决组播报文在二层广播的问题。运用了IGMP Snoopig的设备通过监听主机发出的IGMP报文，判断主机是否期望加入或离开某个组播组，从而建立组播MAC地址表。

IGMP Snooping工作机制

当二层以太网交换机收到主机和路由器之间传递的IGMP报文时，IGMP Snooping分析报文所带的信息：

- 如果主机发出IGMP主机报告报文时，交换机将该主机加入到相应的组播表中。
- 如果主机发出IGMP离开报文时，交换机将删除与该主机对应的组播表项。

通过不断监听IGMP报文，交换机在二层建立和维护组播MAC地址表，交换机根据组播MAC地址表转发从路由器下发的组播报文。

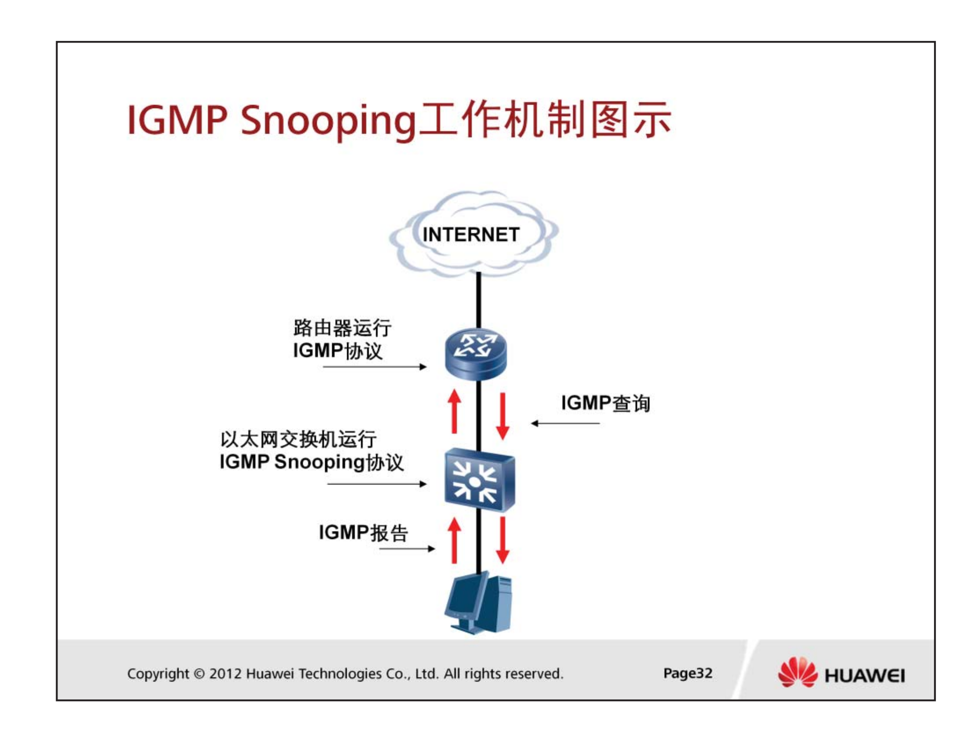
 Page31



IGMP Snooping运行在链路层，是二层以太网交换机上的组播约束机制，用于管理和控制组播组。

当二层以太网交换机收到主机和路由器之间传递的IGMP报文时，IGMP Snooping分析IGMP报文所带的信息。当监听到主机发出的IGMP主机报告报文时，交换机就将该主机加入到相应的组播表中；当监听到主机发出的IGMP离开报文时，交换机就将删除与该主机对应的组播表项。通过不断地监听IGMP报文，交换机就可以在二层建立和维护组播MAC地址表。之后，交换机就可以根据组播MAC地址表转发从路由器下发的组播报文。

没有运行IGMP Snooping时，组播报文将在二层广播，运行IGMP Snooping后，报文将不再在二层广播，而是进行二层组播 。



以太网交换机通过运行IGMP Snooping实现对IGMP报文的侦测，并为主机及其对应端口与相应的组播组地址建立映射关系。为实现IGMP Snooping，二层以太网交换机对各种IGMP报文的处理过程如胶片所示：

当收到IGMP通用查询报文时，如果收到通用查询报文的端口原来就是路由器端口，以太网交换机就重置该端口的老化定时器；如果收到通用查询报文的端口原来不是路由器端口，则交换机启动对该端口的老化定时器。

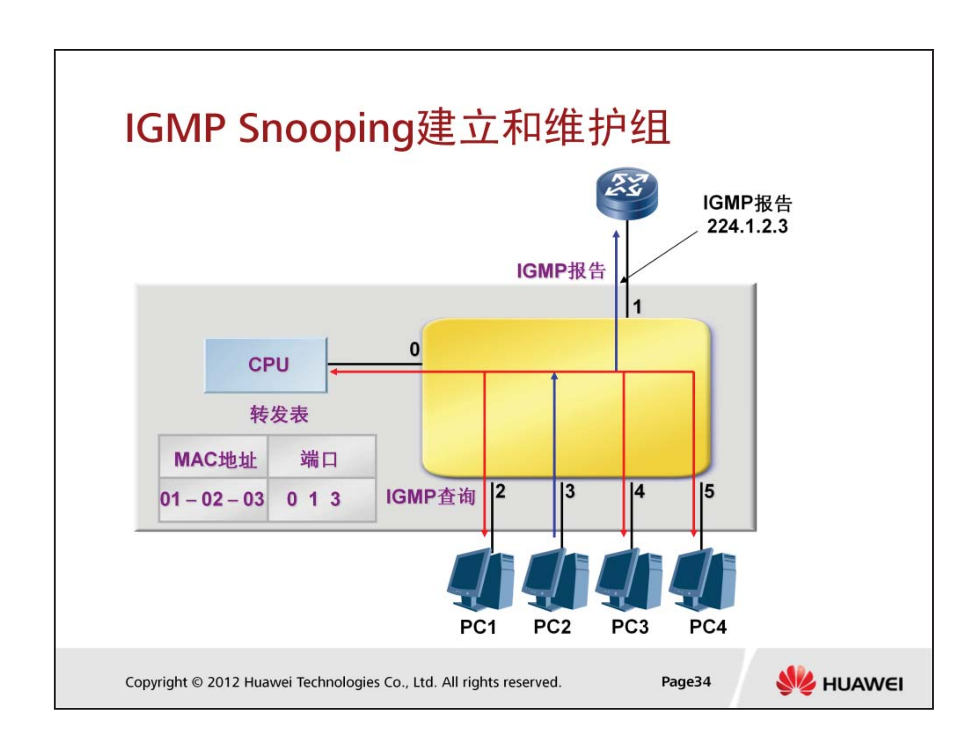
当以太网交换机收到IGMP特定组查询报文时，只向被查询的IP组播组发特定组查询。

当以太网交换机收到IGMP报告报文时，首先判断该报文要加入的IP组播组对应的MAC组播组是否已经存在。 如果不存在则新建MAC组播组，并将接收报告报文的端口加入该MAC组播组中，同时启动该端口的老化定时器，将该端口所属VLAN下存在的所有路由器端口加入到此MAC组播转发表中，而且新建IP组播组，并将接收报告报文的端口加入到IP组播组中。

如果该报文对应的MAC组播组已存在，并且接收报告报文的端口也已经存在于该MAC组播组，则仅重置接收报告报文的端口上的老化定时器。

IGMP离开报文：当以太网交换机收到对某IP组播组的离开报文，则会向

接收此离开报文的端口发送所离开组的特定组查询报文，以确认此端口相连的主机中还有没有此组播组的其他成员，同时启动一个响应查询定时器。如果在该定时器超时的时候还没有收到该组播组的报告报文，则将该端口从相应MAC组播组中删去。如果MAC组播组没有组播成员端口时，交换机将通知组播路由器将该分支从组播树中删除。



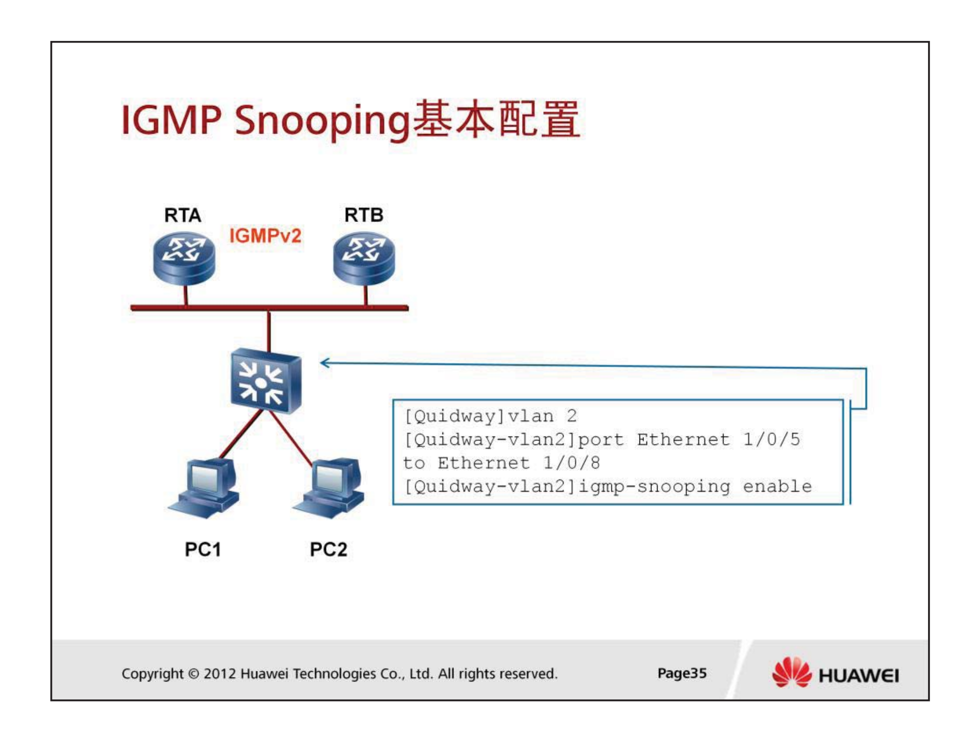
看看利用IGMP Snooping建立和维护组播组的过程：

1、在上图中，PC2希望加入组播组224.1.2.3，因此以组播方式发送一个IGMP成员报告给该组，报告中具有目的MAC地址0x0100.5e01.0203。最初转发表上没有这个组播MAC地址的表项，所以该报告被扩散到交换机的所有端口，包括与交换机CPU相连的内部端口0；

2、当CPU收到PC2组播的IGMP报告时，CPU利用IGMP报告中的信息建立了一个转发表项，此表项包括PC2的端口号，连接路由器的端口号和连接交换机内部CPU的端口号；

3、形成此转发表项的结果是使后面任何目的地址为0x0100.5e01.0203的组播帧都被抑制在端口0、1和3，而且不向交换机其他端口扩散。

4、假设PC4要加入该组，并主动发一个IGMP报告给该组，交换机根据转发表项向外部端口1和3转发此报告。交换机的CPU也收到此报告，它在转发表项上为MAC地址0x0100.5e01.0203增加一个端口（端口5）。



二层交换机上IGMP Snooping的配置相对较简单，在全局或在指定的VLAN视图下使能IGMP Snooping即可。

IGMP Snooping验证

```
[Quidway]display igmp-snooping group
Total 2 IP Group(s).
Total 2 MAC Group(s).
Vlan(id):2.
Total 2 IP Group(s).
Total 2 MAC Group(s).
Router port(s):Ethernet1/0/7     Ethernet1/0/8
IP group(s):the following ip group   IP group address:224.1.1.1
Host port(s):Ethernet1/0/6
    MAC group(s):
        MAC group address:0100-5e01-0101
        Host port(s):Ethernet1/0/6
      ......
```

 Page36 HUAWEI

display igmp-snooping group命令查看交换机上的IGMP信息。包括连接组播路由器的端口，连接主机的端口，组播组IP地址，用户主机的MAC地址。



问 题

IGMPv1提供哪两种类型的报文?

IGMPv2与IGMPv1相比较增加了哪些功能?

IGMP Snooping的原理与作用?

Page37



1、IGMPv1提供哪两种类型的报文?

IGMPv1提供成员关系查询和成员关系报告两种类型的报文。

2、IGMPv2与IGMPv1相比较增加了哪些功能?

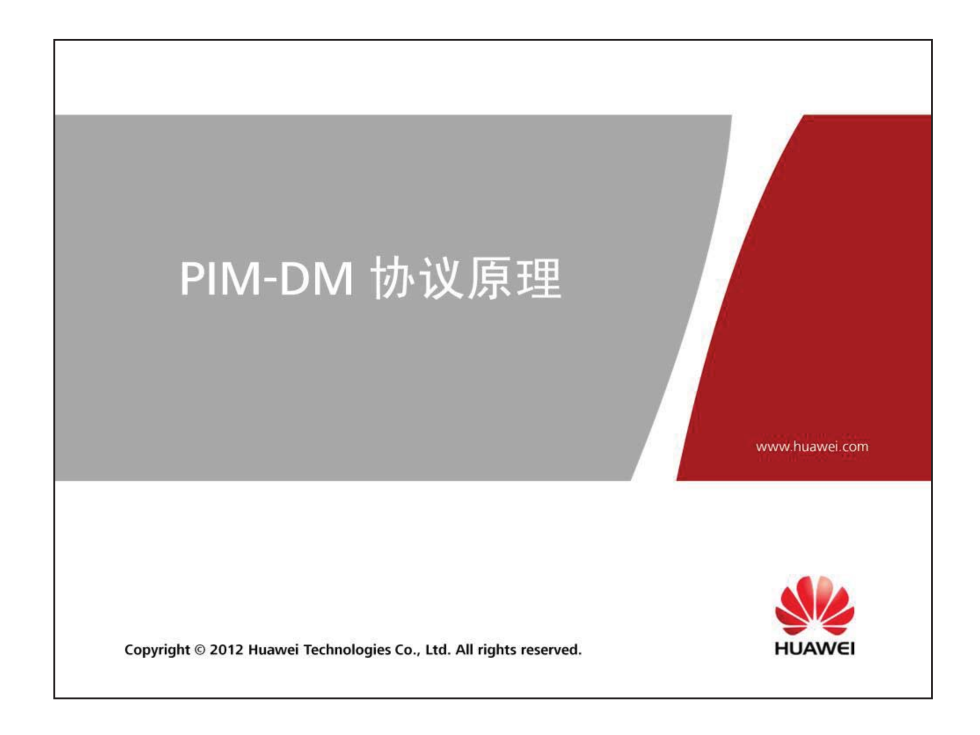
IGMPv2中增加了组播成员离开功能，当组播成员要离开组播组时会发送一个离开报文给组播组。同时IGMPv2在增加了特定组查询报文，用于成员要离开组播组时查询组内是否还有其他成员。

3、IGMP Snooping的原理?

IGMP Snooping运行在链路层，是二层以太网交换机上的组播约束机制，用于管理和控制组播组。

二层以太网交换机收到主机和路由器之间传递的IGMP报文时，IGMP Snooping分析IGMP报文所带的信息，不断地监听IGMP报文，交换机在二层建立和维护组播MAC地址表，交换机根据组播MAC地址表转发从路由器下发的组播报文。


谢谢
www.huawei.com


PIM-DM 协议原理
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

组播路由器之间运行组播路由协议，组播路由协议用于建立和维护组播路由，并正确、高效地转发组播数据包。

PIM（Protocol Independent Multicast）是典型的域内组播路由协议，分为DM（Dense Mode）和SM（Sparse Mode）两种模型。

Page1



与组播相关的协议中前面已经介绍到组播组管理协议IGMP。组播组管理协议IGMP（Internet Group Management Protocol）在接收者主机和组播路由器之间运行。路由器之间则需要运行组播路由协议。

组播路由协议用于建立和维护组播路由，并正确、高效地转发组播数据包。组播路由形成了一个从数据源到多个接收端的单向无环数据传输路径，即组播分发树。组播路由协议分为域内组播路由和域间组播路由协议。这里介绍最典型的域内组播路由协议PIM。



培训目标

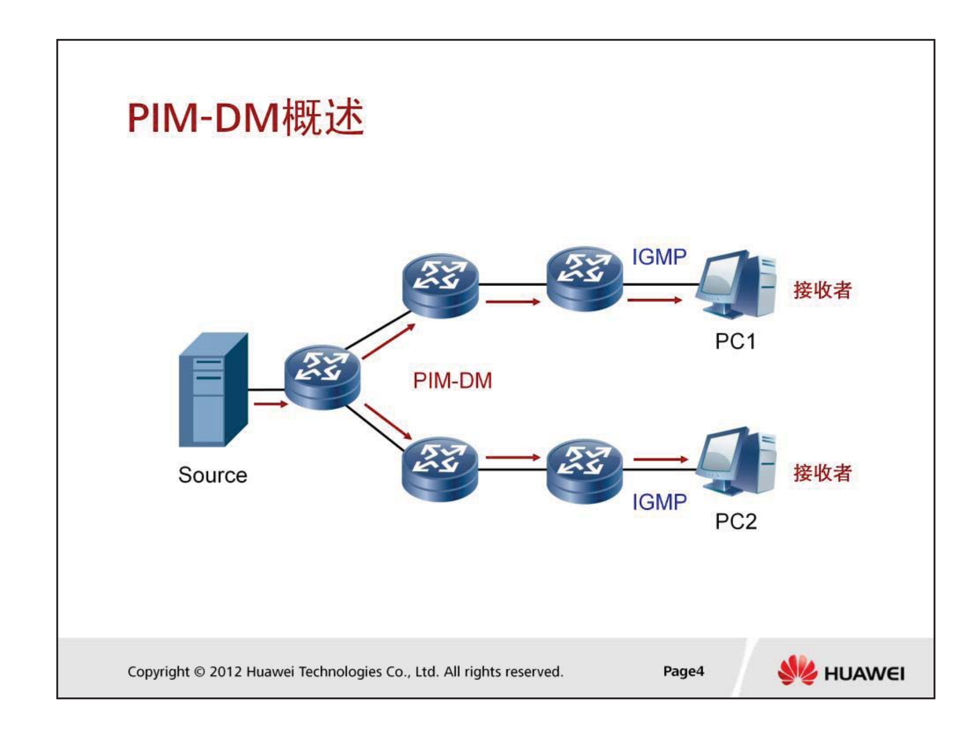
学完本课程后，您应该能：

- 理解PIM-DM协议基本原理
- 理解PIM-DM协议工作机制
- 掌握PIM-DM的基本配置





本章给出PIM-DM协议的基本原理。并列举一个配置实例使学员可以对PIM-DM有一个总体的认识。



PIM（Protocol Independent Multicast）称为协议无关组播，即给IP组播提供路由信息的可以是静态路由、RIP、OSPF、IS-IS、BGP等任何一种单播路由协议。组播路由和单播路由协议无关，只要通过单播路由协议能够产生相应组播路由表项即可。

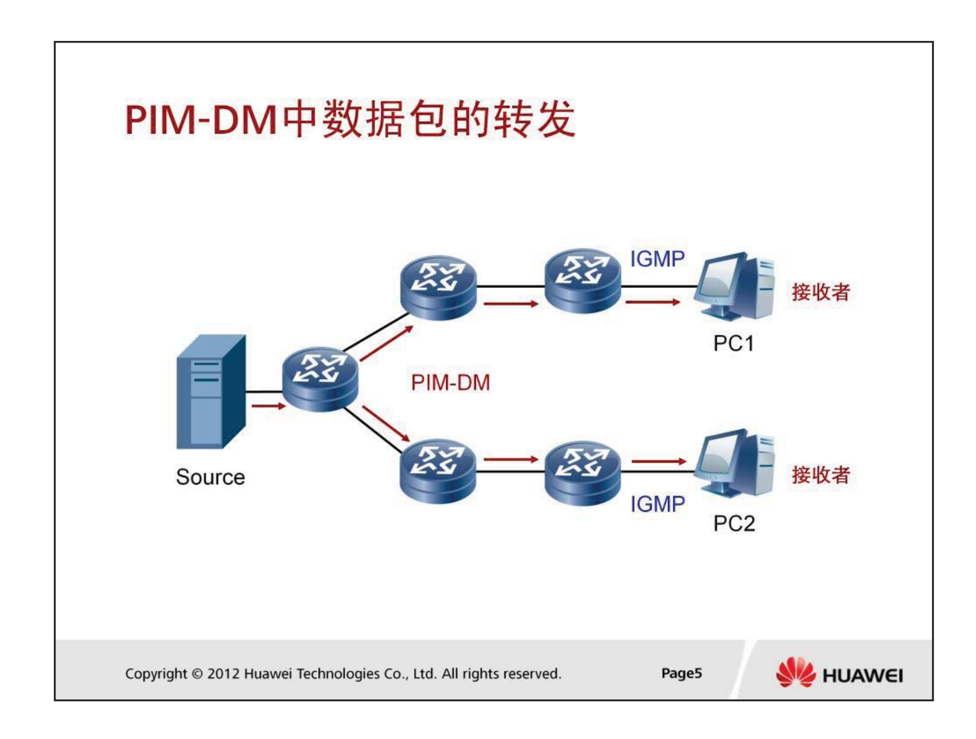
PIM-DM（Protocol Independent Multicast Dense Mode）称为协议独立组播－密集模式，属于密集模式的组播路由协议，适用于小型网络。在这种网络环境下，组播组的成员相对比较密集。

PIM-DM假设网络中的每个子网都存在至少一个对组播源感兴趣的接收站点，因此组播数据包被扩散到网络中的所有点，与此伴随着相关资源（带宽和路由器的CPU等）的消耗。

为了减少网络资源的消耗，密集模式组播路由协议对没有组播数据转发的分支进行Prune操作，只保留包含接收者的分支。

被剪掉的分支如果有组播数据转发需求也可以重新接收组播数据流。PIM-DM使用Graft嫁接机制主动恢复组播报文的转发。

周期性的扩散和剪枝现象是密集模式协议的特征。



DM模式下数据包的转发路径是一颗“有源树”。

“有源树”是以“组播源”为根、组播组成员为枝叶的一棵树。有源树使用的是从组播源到接收者的最短路径，因此也称为最短路径树SPT（Shortest Path Tree）。如图中箭头所示的从Source到接收者的路径。

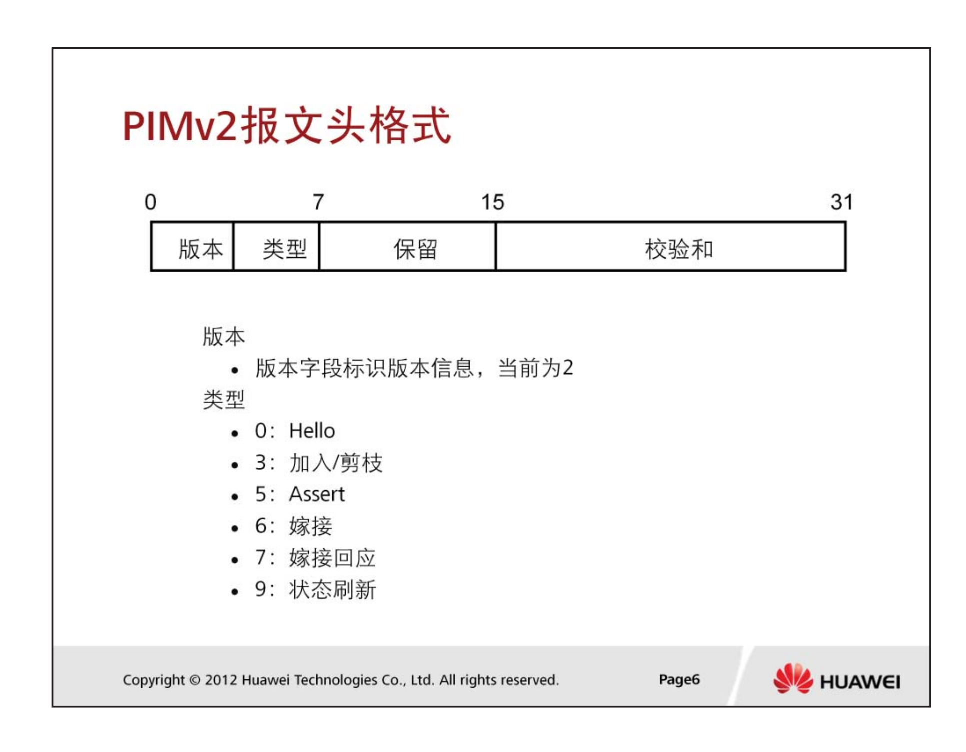
PIM-DM不依赖于特定的单播路由协议，而是使用现存的单播路由表进行RPF检查。

数据包的转发中会出现上游和下游两个概念。

路由器收到组播数据的接口称为上游。

转发组播数据的接口称为下游。

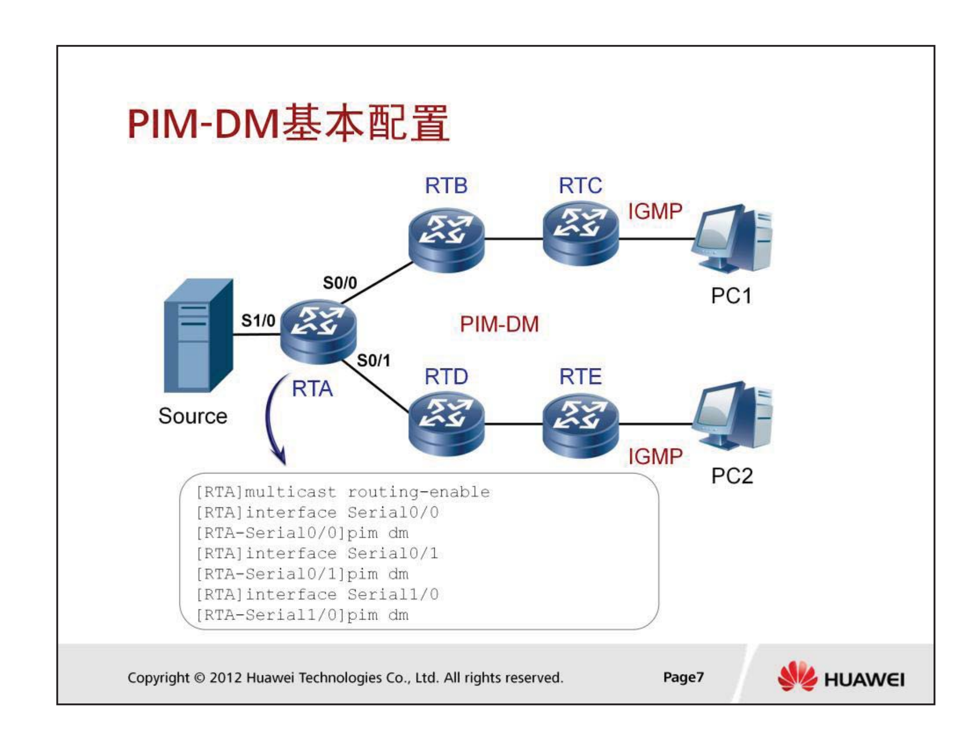
数据包的转发是从上游至下游方向的转发。



PIM协议报文中的协议号都是103。

本页列出的几种报文类型是于PIM-DM协议相关的，PIM-SM协议类型的报文将在PIM-SM课程中介绍。

PIM-DM协议有报文类型有：Hello报文、加入/剪枝消息、Assert报文（Assert消息使用组播方式发送，目的地是224.0.0.13的所有PIM路由器）、嫁接消息、嫁接回应消息，状态刷新。这几类报文主要用于周期的建立、维护SPT树。在PIM-DM协议工作机制中会详细介绍以上几类报文的作用。



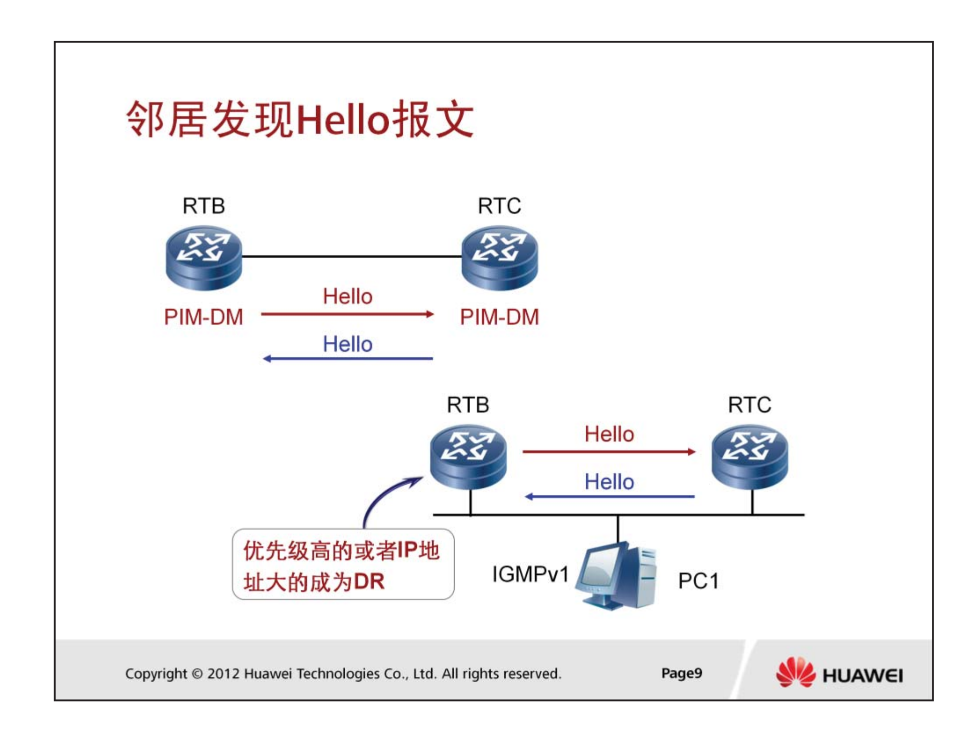
做为组播路由协议PIM-DM的配置十分简单。只需在路由器的接口上使能PIM-DM协议即可。

注意：在接口下使能PIM-DM之前必须首先全局使能IP组播路由。

命令为：multicast routing-enable



对PIM-DM有了一个总体的认识后，本章详细介绍PIM-DM协议机制。即“扩散—剪枝—嫁接”过程。



在PIM-DM网络中，刚启动的组播路由器需要使用Hello消息来发现邻居，并维护邻居关系。路由器之间周期性地发送Hello消息来构建和维护SPT树。

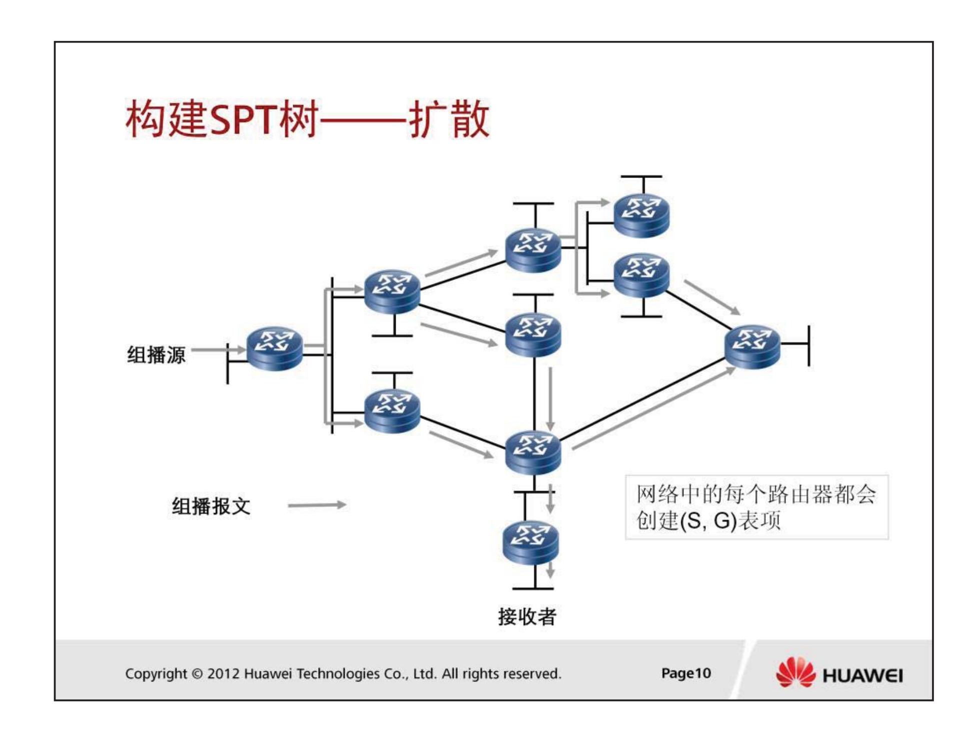
pim timer hello interval，在接口视图下配置发送Hello消息的时间间隔。Hello消息默认周期是30秒。

除了维护邻居关系外，Hello消息还具有一个重要的功能就是在多路由器网段中选举DR指定路由器。DR充当IGMPv1查询器。在IGMP概述中已经提到过IGMPv1中查询器的选举由组播路由协议决定。

在PIM-DM中各路由器通过比较Hello消息上携带的优先级和IP地址，为多路由器网段选举指定路由器DR，充当IGMPv1的查询器。

当DR出现故障时，接收Hello消息将会超时，邻居路由器之间会触发新的DR选举过程。

pim hello-option holdtime interval 在接口视图下配置Hello消息超时时间值。默认情况超时时间值为105秒。

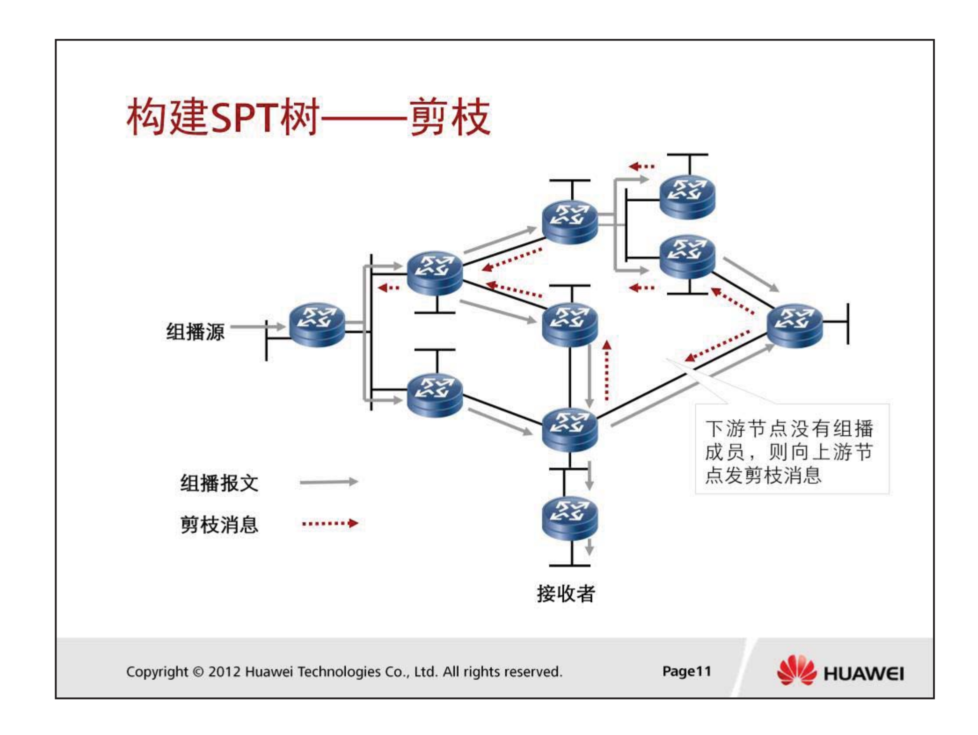


PIM-DM假设网络上的所有主机都准备接收组播数据，当某组播源S开始向组播组G发送数据时，具体过程如下：

路由器接收到组播报文后，首先根据单播路由表进行RPF检查。

- 如果检查通过则创建一个（S，G）表项，然后将数据向网络上所有下游PIM-DM节点转发，这个过程称为扩散（Flooding）。
- 如果没有通过RPF（Reverse Path Forwarding）检查，即组播报文从错误的接口接收，则将报文丢弃。

经过这个过程，PIM-DM组播域内每个路由器上都会创建（S，G）表项。

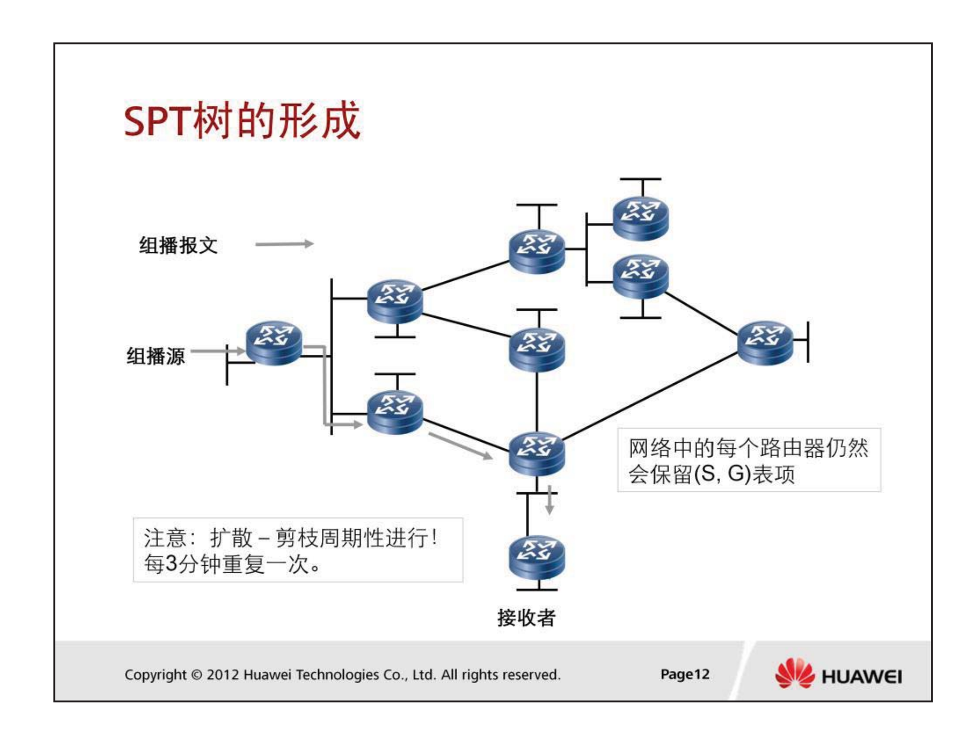


无论下游有没有组播成员，组播报文都会被扩散出去，因此会导致带宽资源的浪费。为避免带宽的浪费PIM-DM使用剪枝机制。

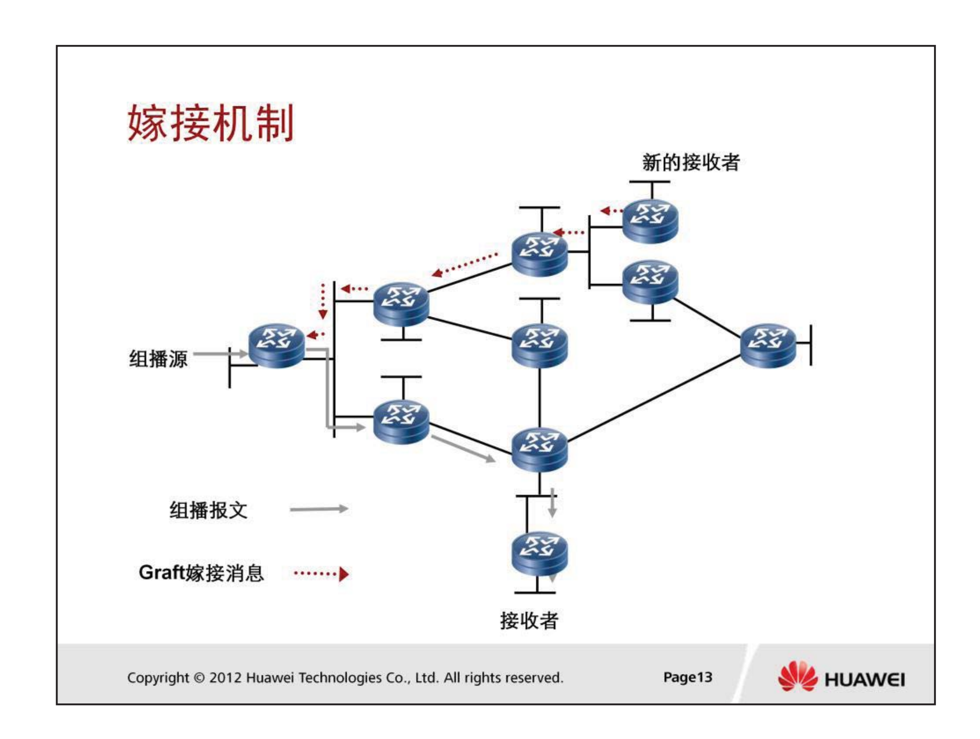
当下游节点没有组播组成员，则路由器向上游节点发Prune剪枝消息，通知上游节点不用再转发数据到该分支。上游节点收到Prune剪枝消息后，就将相应的接口从其组播转发表项（S，G）对应的输出发送列表中删除。剪枝过程继续直到PIM-DM中仅剩下了必要的分支，这就建立了一个以组播源S为根的SPT（一种组播转发树，被称为：源路径树或最短路径树）。

各个被剪枝的节点同时提供超时机制，当剪枝超时时重新开始扩散—剪枝过程。剪枝状态超时计时器的默认值为210秒。

PIM-DM的扩散—剪枝机制周期性进行。

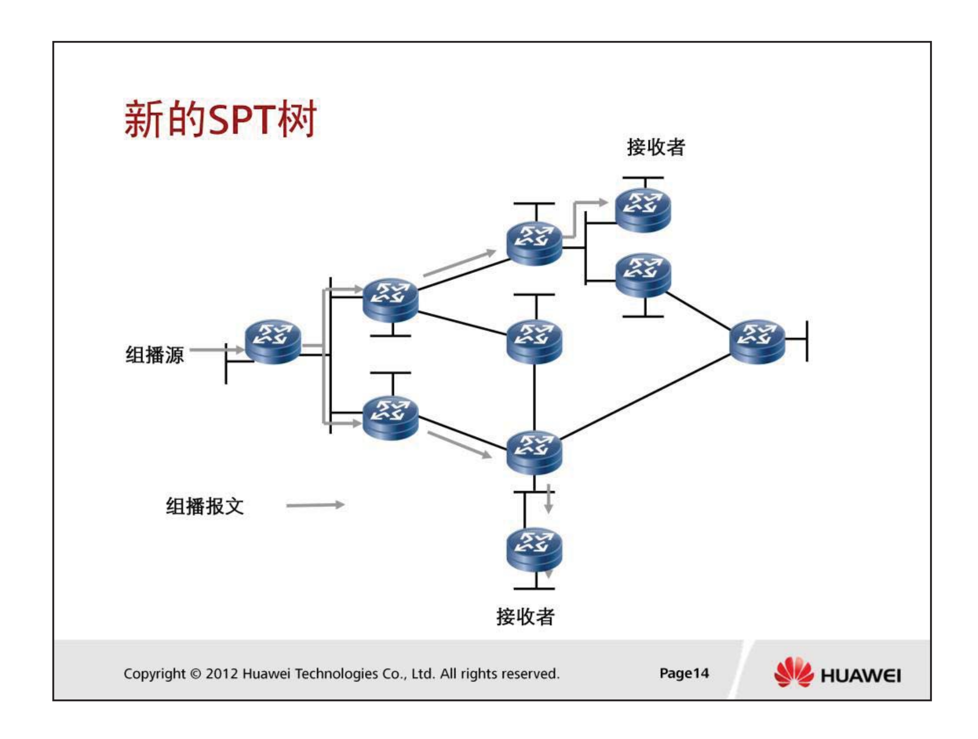


组播路由器根据剪树消息剪去多余的分枝，形成一棵新的SPT树。虽然剪枝消息让路由器不再向没有组播成员的分枝转发组播报文。但是每个路由器上的（S，G）表项仍存在，其目的是为了一旦有组播成员加入时可以快速加入并转发组播报文。

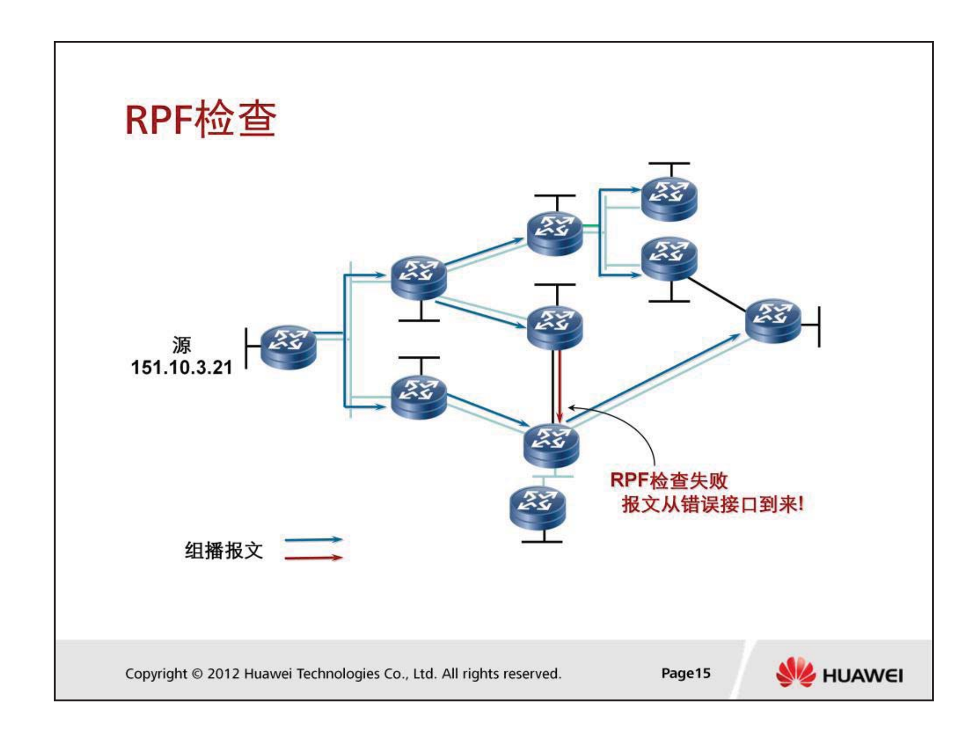


被剪枝的下游节点在剪枝超时计时器超时时可以恢复到转发状态，但是剪枝超时计时器要等待210秒。如果在这期间有组播成员想加入则必须等待，这个时间是比较长的。为了减少反应的时间，当被剪枝的下游节点需要恢复到转发状态时，该节点可以使用Graft嫁接消息主动通知上游节点。

如上图所示：网络中一个接收者恢复接收组播数据，Graft嫁接消息逐跳向组播源S传递，中间节点接收到Graft嫁接消息后回应确认，从而先前被剪掉的分支恢复信息传输。



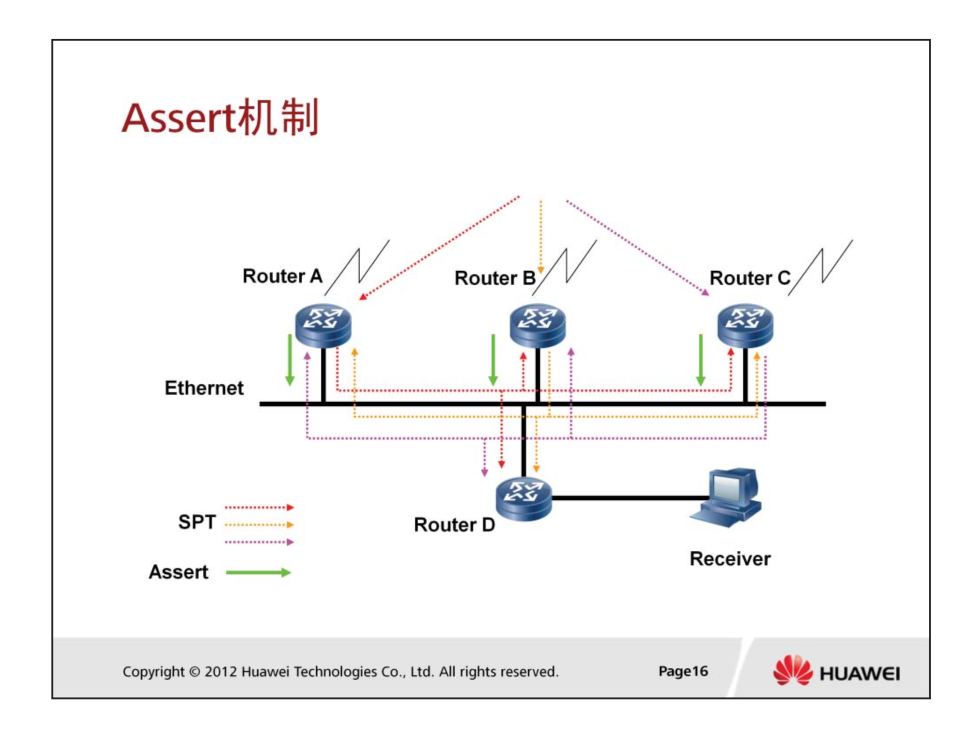
剪枝路径重新恢复为转发路径，生成一棵新的SPT树。



PIM-DM采用RPF检查机制，利用现存的单播路由表、组播静态路由表、MBGP路由表来构建一棵从数据源S始发的组播转发树。

当一个组播包到达时，路由器首先判断到达路径的正确性。如果到达接口是单播路由指示的通往组播源S的接口，就认为这个组播包是从正确路径而来；否则，将组播包作为冗余报文丢弃。

作为路径判断依据之一的单播路由信息可以来源于任何一种单播路由协议，如RIP、OSPF发现的路由信息，不依赖于特定的单播路由协议。



在共享网络（如Ethernet）中会出现相同报文的重复发送。如上图所示：LAN网段上包含多台组播路由器A、B、C和D，各自都有到组播源S的接收途径。当路由器A、B和C都从上游接收到组播源发出的组播数据报文后，都会向Ethernet网络上转发该组播报文，这时下游节点组播路由器D就会收到三份完全相同的组播报文。

为了避免这种情况，就需要通过Assert机制来选定一个唯一的转发者。网络中的各路由器通过发送Assert报文选出一条最优的路径。

选举机制如下：

如果两条或两条以上路径的优先级和到组播源的开销相同，则IP地址最大的路由器获胜成为该（S，G）项的上游邻居，由它负责该（S，G）组播报文的转发，而其他落选路由器则剪掉对应的接口以禁止转发信息。

PIM-DM协议机制

PIM-DM的工作过程可以概括为：

- 邻居发现
- 扩散
- 剪枝
- 嫁接
- Assert机制

 Page17 

PIM-DM的工作过程可以概括为：邻居发现、扩散、剪枝、嫁接阶段、Assert机制。

邻居发现

组播路由器使用Hello消息发现并维护邻居关系。并通过比较Hello消息上携带的优先级和IP地址，各路由器为多路由器网段选举指定路由器DR，充当IGMPv1的查询器。

扩散（Flooding）

组播源S向组播组G发送数据时，路由器接收到组播报文后，首先根据单播路由表进行RPF检查，通过则创建一个（S，G）表项，然后将数据向网络上所有下游PIM-DM节点转发，这个过程称为扩散（Flooding）。没有通过RPF检查，则将报文丢弃。

剪枝（Prune）

如果下游节点没有组播组成员，则向上游节点发Prune剪枝消息，通知上游节点不用再转发数据到该分支。上游节点收到Prune剪枝消息后，就将相应的接口从其组播转发表项（S，G）对应的出接口列表中删除。剪枝过程继续直到PIM-DM中仅剩下了必要的分支，建立了一个以组播源S为根的SPT（一种组播转发树，被称为：源分发树或最短路径树）。

嫁接（Graft）

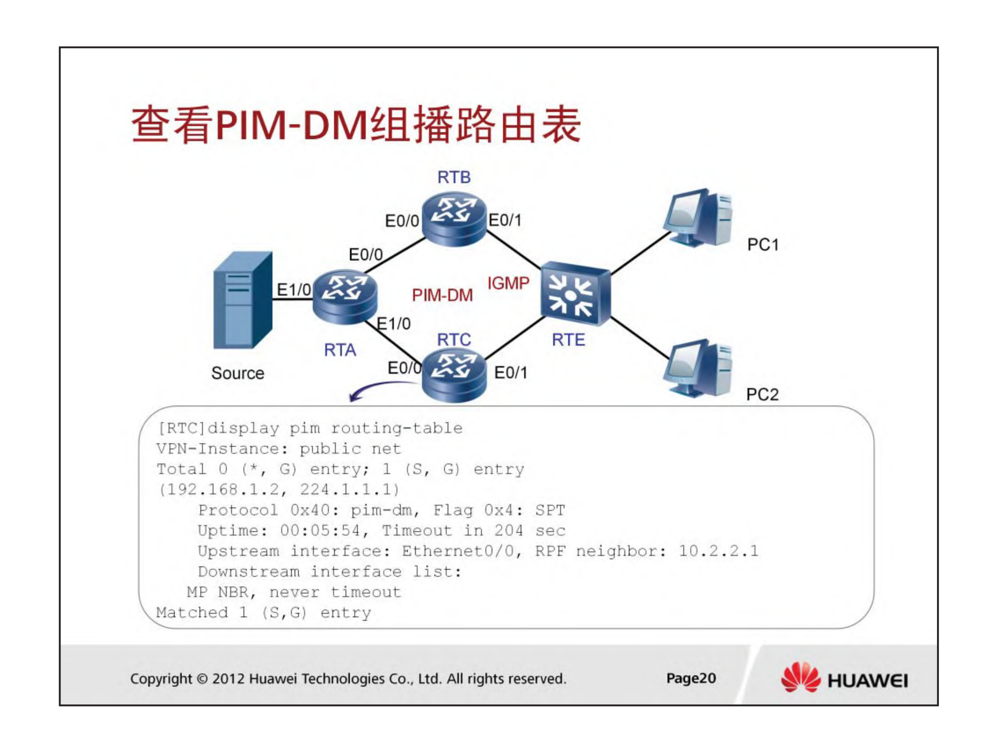
当被剪枝的下游节点需要恢复到转发状态时，该节点使用Graft嫁接消息通知上游节点恢复信息传输。

Assert机制

在共享网络使用Assert机制指定转发器。

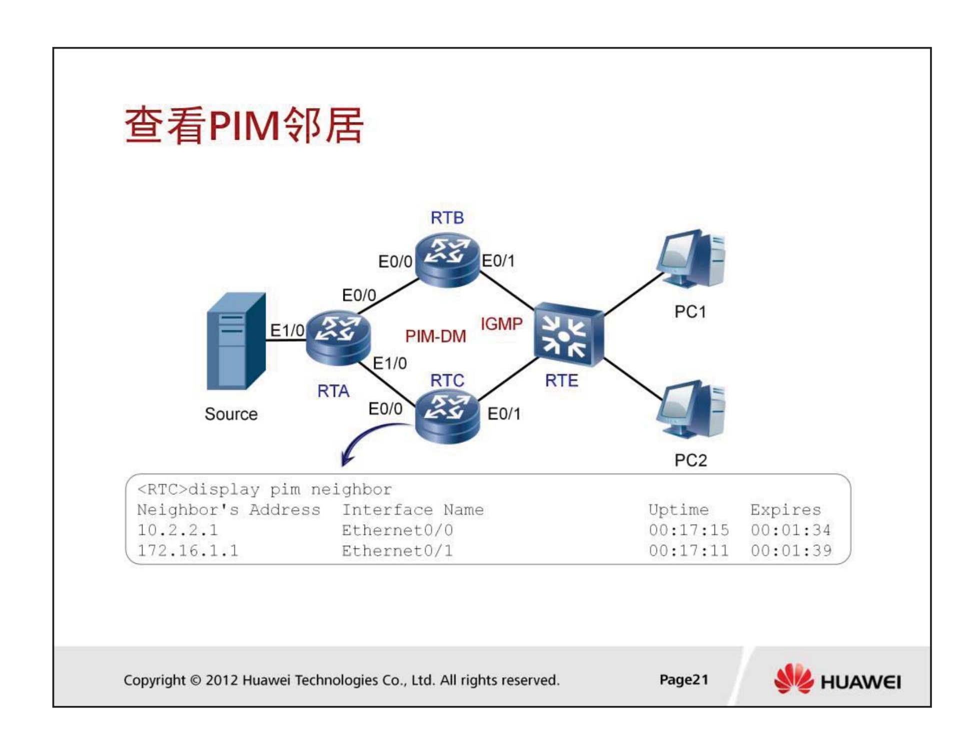


本章介绍如何验证PIM-DM配置。



display pim routing-table 查看路由器PIM协议组播路由表。路由器接收到组播报文后后生成（S，G）表项，（S，G）表项中会列出组播报文所经过的路由器中的节点信息。即上游节点和下游节点，以及RPF邻居关系。RPF邻居信息用于RPF检查判断报文是否从正确接口处接收。

如果某一路由器的组播路由表为空则需要检查相应的配置。



display pim neighbor查看路由器PIM邻居信息。

正常情况下RTC有两个邻居，分别是RTA（10.2.2.1)和RTB(172.16.1.1)。

查看PIM接口配置

```
[RTC]display pim interface verbose
PIM information of interface Ethernet0/0:
  IP address of the interface is 10.2.2.2
  PIM is enabled
  PIM version is 2
  PIM mode is Dense
  PIM query interval is 30 seconds
  PIM neighbor limit is 128
  PIM neighbor policy is none
  Total 1 PIR(designated router) is 10.2.2.2
PIM information of interface Ethernet0/1:
  IP address of the interface is 172.16.1.3
  PIM is enabled
  PIM version is 2
  PIM mode is Dense
  PIM query interval is 30 seconds
  PIM neighbor limit is 1e
  Total 1 PIM neighbor on interface
  PIM DR(designated router) is 172.16.1.3
```

 Page22 HUAWEI

display pim interface verbose命令用来查看接口上的PIM详细信息。包括是否使能PIM协议，PIM的模式以及DR路由器IP地址信息。

display pim interface命令用来查看接口上PIM的简略信息；



问 题

PIM-DM的基本原理?

PIM-DM中嫁接的作用?

PIM-DM中Assert机制的作用?

 Page23



1、PIM-DM的基本原理?

PIM-DM假设网络中的每个子网都存在至少一个对组播源感兴趣的接收站点，组播数据包扩散到网络中的所有点。对没有组播数据转发的分支PIM-DM进行Prune剪枝操作，只保留包含接收者的分支。使用Graft嫁接机制剪掉的分支可以恢复成转发状态。

2、PIM-DM中嫁接的作用?

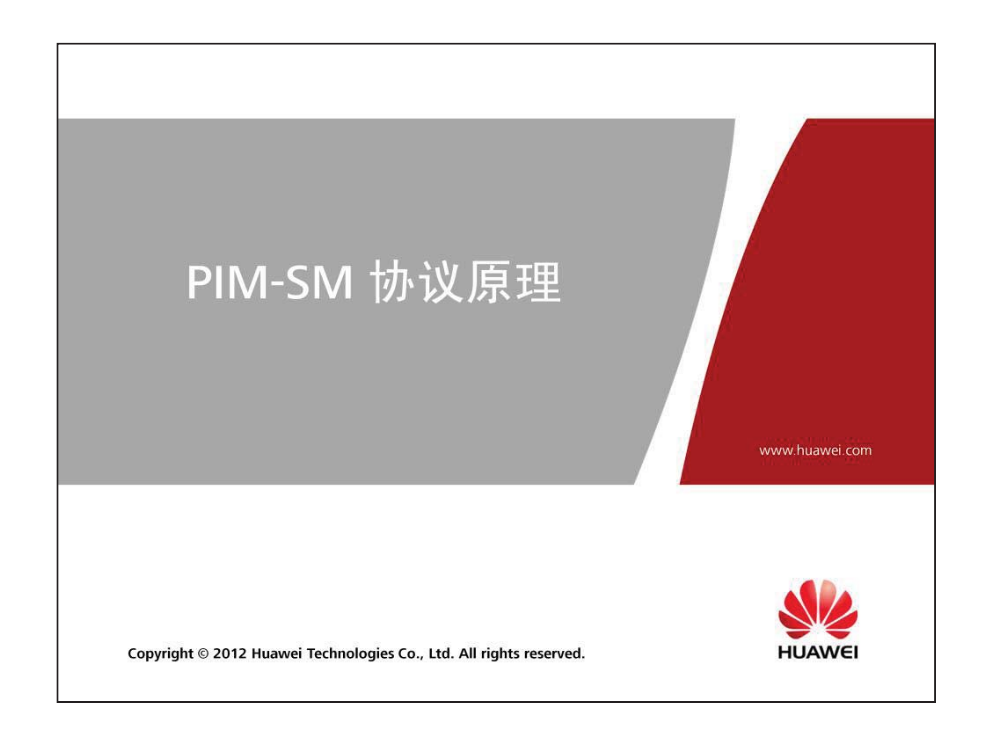
路由器由剪枝消息剪去多余的没有组播成员的分支，而被剪枝处如果有组播成员要加入，则需要等待超时器。使用嫁接机制，当被剪枝的下游节点需要恢复到转发状态时，该节点可以主动通知上游节点有组播成员加入，从而减少响应时间。

如上所示：网络中一个接收者恢复接收组播数据，Graft嫁接消息逐跳向组播源S传递，中间节点接收到Graft嫁接消息后回应确认，从而先前被剪掉的分支恢复信息传输。

3、PIM-DM中Assert机制的作用?

Assert可以避免在共享网络（如Ethernet）中相同报文的重复发送。通过Assert机制在共享网络中来选定一个唯一的转发者。其他落选路由器则剪掉对应的接口以禁止转发信息。


谢谢
www.huawei.com


PIM-SM 协议原理
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

PIM-SM(Protocol Independent Multicast-Sparse Mode)称为协议无关组播－稀疏模式。

属于稀疏模式的组播路由协议，适用于组成员分布相对分散、范围较广、大规模的网络。

 Page1



PIM-DM协议假定所有节点都有组播成员，因此会将组播报文扩散到所有下游分支。虽然有剪枝机制可以剪去没有组播成员的分支，但由于这种“扩散—剪枝”不够高效，再加上PIM-DM使用的是SPT树，不支持共享树，因此只适合组成员分布相对密集的小型网络。

PIM-SM（Protocol Independent Multicast-Sparse Mode）称为协议无关组播－稀疏模式，适用于组成员分布相对分散、范围较广、大规模的网络。



培训目标

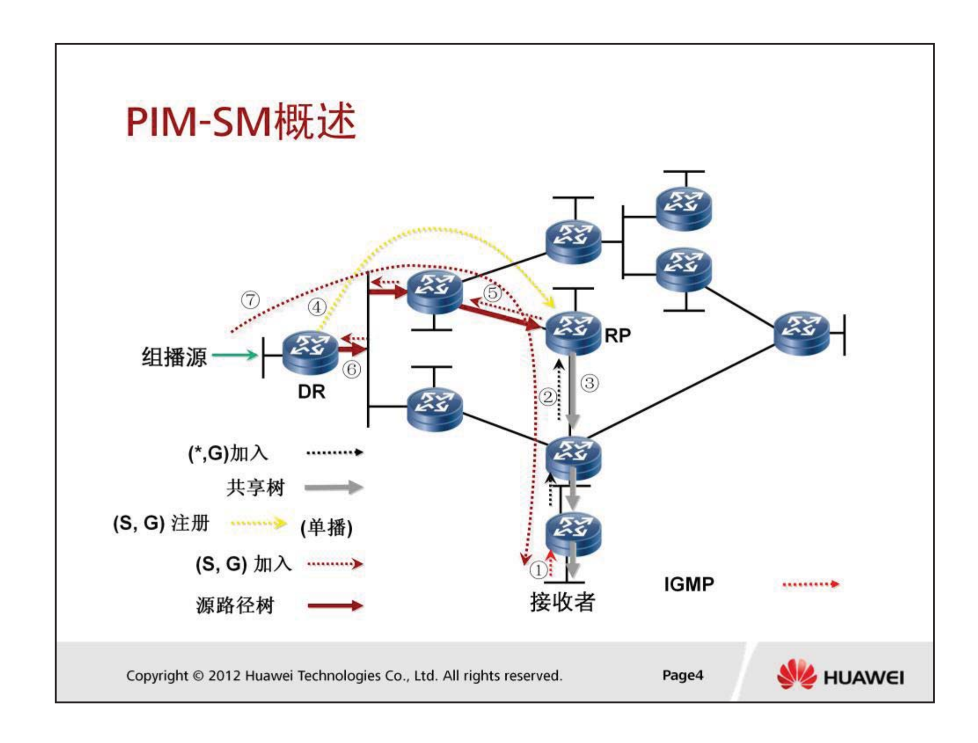
学完本课程后，您应该能：

- 掌握PIM-SM的基本原理和配置
- 掌握共享树的加入和源的注册过程
- 掌握RPT向SPT的切换





本章简单介绍PIM-SM协议原理，并介绍PIM-SM的基本配置。



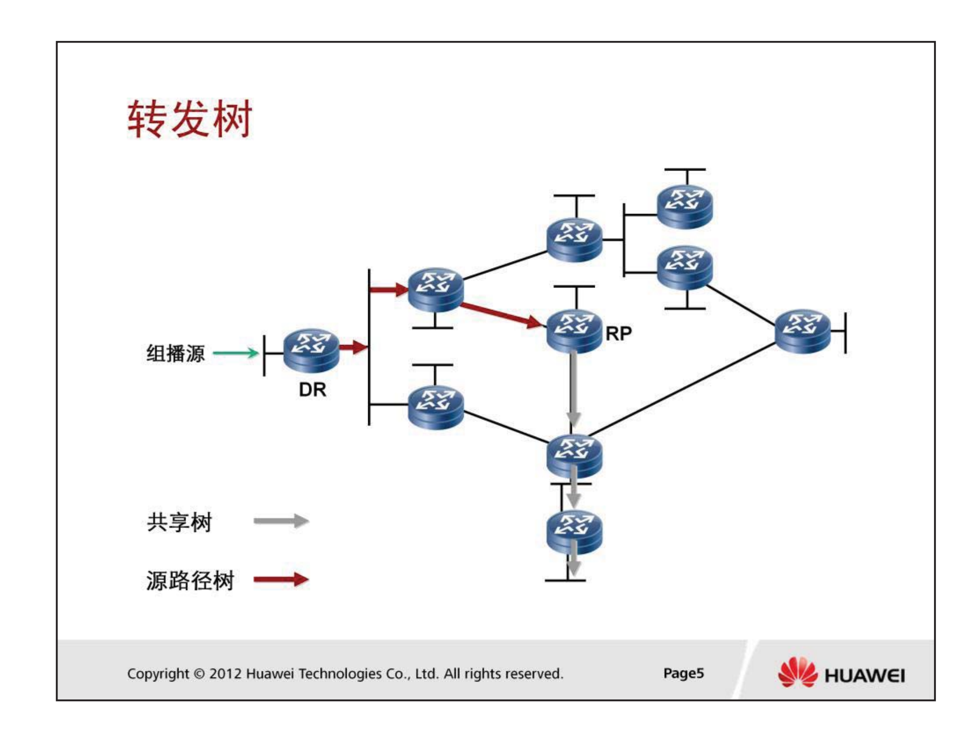
PIM-SM假设网络中的组成员分布非常稀疏，几乎所有网段均不存在组成员。直到某网段出现组成员时，才构建组播路由，向该网段转发组播数据。

PIM-SM模型实现组播转发的核心任务是构造并维护一棵单向共享树。

共享树选择PIM中某一路由器作为公用根节点，称为汇聚点RP（Rendezvous Point）。组播数据通过RP沿共享树向接收者转发。

接收侧，连接信息接收者的路由器向该组播组对应的RP发送组加入消息，加入消息经过一个个路由器后到达根部（即RP汇聚点），所经过的路径就变成了此共享树RPT的分支。

发送端如果想要往某组播组发送数据，首先由第一跳路由器向RP汇聚点进行注册，注册消息到达RP后触发源树建立。之后组播源把数据发向RP汇聚点，当数据到达了RP汇聚点后，组播数据包被复制并沿着RPT树传给接收者。



PIM-SM同时包含两种树：共享树和源路径树。

从RP到组播接收者数据转发的路径称为共享树。

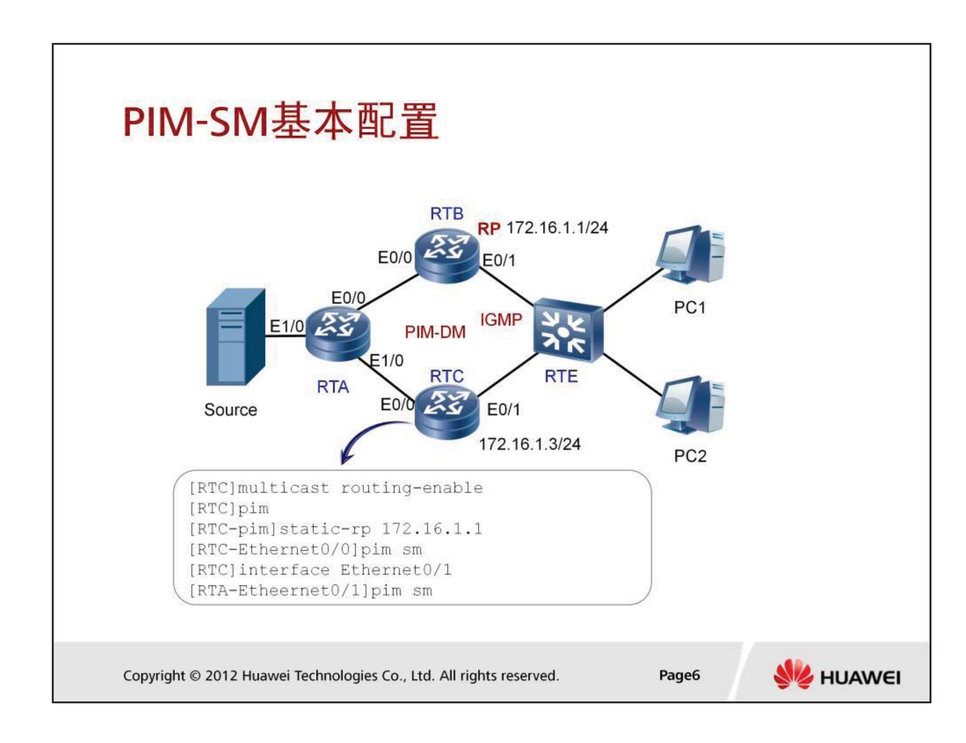
从组播源到RP的数据转发路径称为源路径树。

RPF检查根据树的种类进行：

在共享树下，使用RP地址作为检测地址。

在源路径树下，使用组播源地址作为检测地址。

RPF检查机制将在第二章中详细介绍。



PIM-SM的基本配置同PIM-DM一样，都需要使能组播路由协议和在路由器的接口下使能PIM-SM协议。

通过对PIM-SM原理的简单介绍，相信学员已经知道在PIM-SM网络有一个重要的角色就是RP。RP是做为RPT共享树的根，负责组播报文的转发，组播源的加入等。

RP发现有多种方式，在第二章中会详细介绍RP的发现和选举。这里以手工指定RP为RTB(172.16.1.1)为例，简单介绍PIM-SM的基本配置。

基本的PIM-SM配置比较简单。只需要三个步骤：

1、全局使能组播路由协议；

2、接口下使能PIM-SM协议；

3、指定RP地址。

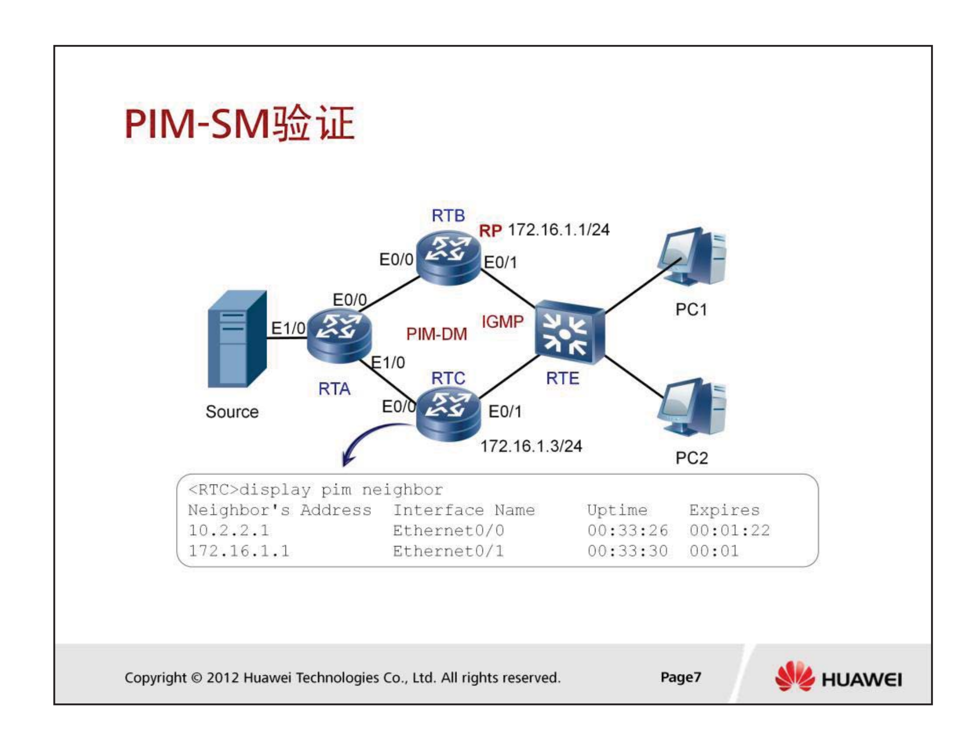
需要注意的是：如果RP是手工指定的，则必须在每一台路由器中都手工配置RP的地址。命令如下：

进入PIM视图

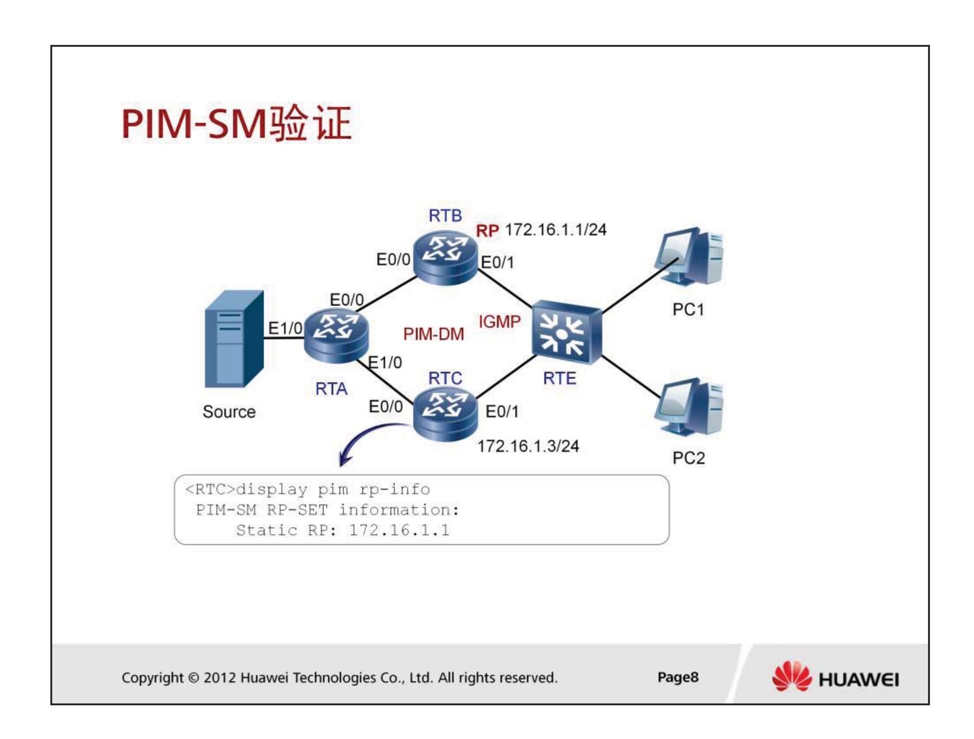
[RTC]pim

手动指定RP地址为172.16.1.1

[RTC-pim]static-rp 172.16.1.1



Display pim neighbor 查看PIM-SM邻居信息。配置成功后RTC建立了两个PIM邻居，分别是RTA和RTB。



查看PIM-SM网络中RP信息。

本例中RP是手工指定的，RP前有Static，说明是手工指定的。

PIM-SM验证

```
<RTC>display pim interface verbose
PIM information of interface Ethernet0/0:
  IP address of the interface is 10.2.2.2
  PIM is enabled
  PIM version is 2
  PIM mode is Sparse
  PIM query interval is 30 seconds
  PIM neighbor limit is 128
  PIM neighbor policy is none
  Total 1 PIM neighbor on interface
  PIM DR(designated router) is 10.2.2.2
PIM information of interface Ethernet0/1:
  IP address of the interface is 172.16.1.3
  PIM is enabled
  PIM version is 2
  PIM mode is Sparse
  PIM query interval is 30 seconds
  PIM neighbor lcy is none
  Total 1 PIM neighbor on interface
  PIM DR(designated router) is 172.16.1.3
```


HUAWEI

Display pim interface 命令查看PIM接口信息。包括是否使能PIM协议，PIM的版本，PIM的模式，PIM查询周期间隔，DR路由器IP地址。

PIM-SM验证

```
<RTB>display pim routing-table
VPN-Instance: public net
Total 1 (*, G) entry; 1 (S, G) entry
(*, 224.1.1.1), RP 172.16.1.1
    Protocol 0x20: pim-sm, Flag 0x2003: RPT WC NULL_IIF
    Uptime: 00:22:48, Timeout in 161 sec
    Upstream interface: Null, RPF neigterface list:
      Ethernet0/1, Protocol 0x100: RPT, timeout in 161 sec
(192.168.1.2, 224.1.1.1)
    Protocol 0x20: pim-sm, Flag 0x80004: SPT
    Uptime: 00:22:44, Timeout in 200 sec
    Upstream interface: Ethernet0/0, RPF neighbor:ace list: NULL
```

 Page10 

组播路由表收录所有PIM路由表项，并下刷到转发表中，由转发表项直接指导组播报文转发。

PIM中存在两种转发表项：(S,G)或(*,G)。S表示组播源，G表示组播组，*表示任意。

(S,G)只适用于源地址为S，组地址为G的组播报文。通常将源地址为S，组地址为G的组播报文表示为(S,G)报文。

(*,G)适用于组地址为G的组播报文。即：不论是哪个组播源发出的，只要是发往组播组G的组播报文，都应该从（*，G）表项中的下游接口转发出去。

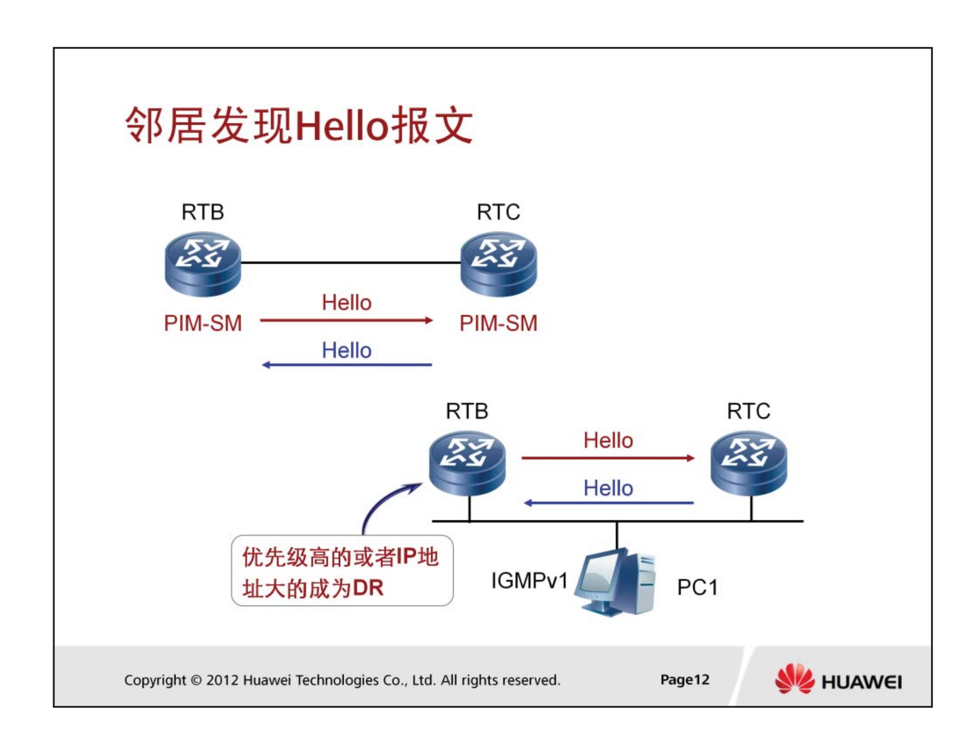
RTB上使用display pim routing-table看到两个表项，(S,G)表项和(*,G)表项。

(S,G)在组播源到RP的SPT树上的路由器会建立该表项。

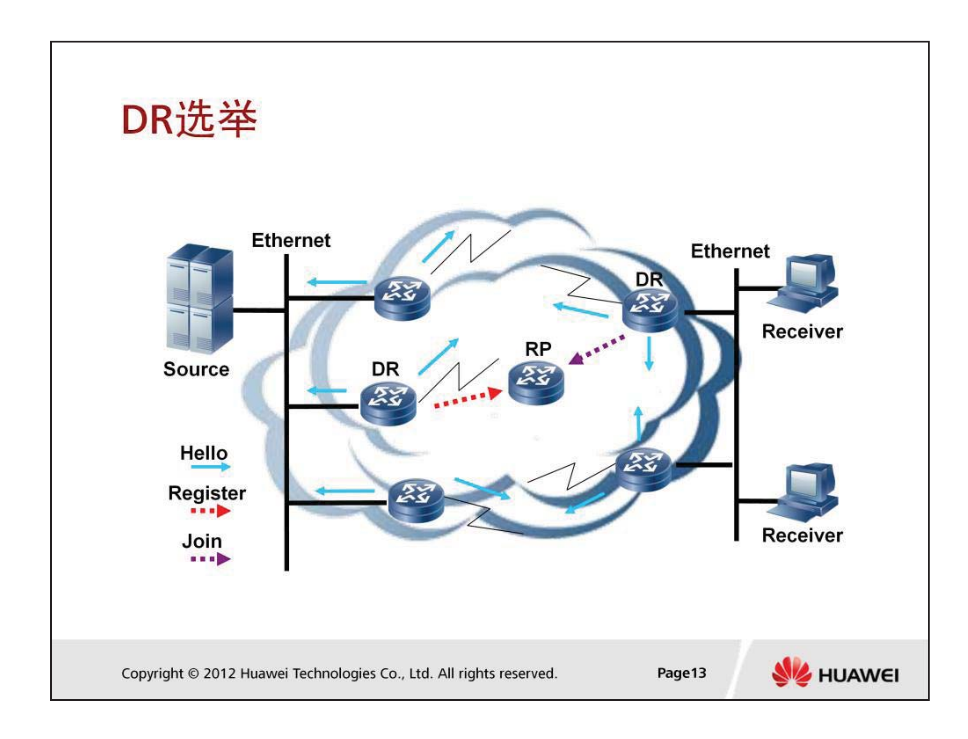
(*,G)则是PIM-SM才有的表项，表示（任意组播源，组播组），在RPT共享树上的路由器会建立该表项。



本章介绍PIM-SM协议工作机制。详细介绍了PIM-SM的工作过程，包括邻居发现、DR选举、RP发现、加入、剪枝、注册、SPT切换。



在PIM-SM网络中，刚启动的组播路由器需要使用Hello消息来发现邻居，并维护邻居关系。通过各路由器之间周期性地使用Hello消息保持联系。除了维护邻居关系外，Hello消息还具有一个重要的功能就是在多路由器网段中选举DR指定路由器。DR充当IGMPv1查询器。



PIM-SM在共享网络（如Ethernet）同样选举DR(Designated Router)。

DR（Designated Router）应用在PIM-SM网络中的如下两个位置：

在连接组播源的共享网段，由DR负责向RP发送Register注册消息。与组播源相连的DR称为源端DR。

在连接组成员的共享网段，由DR负责向RP发送Join加入消息。与组成员相连的DR称为组成员端DR。

共享媒介网络上的各路由器相互之间发送Hello消息（携带DR优先级选项），拥有最高DR优先级路由器将被选举为本网络中的DR。

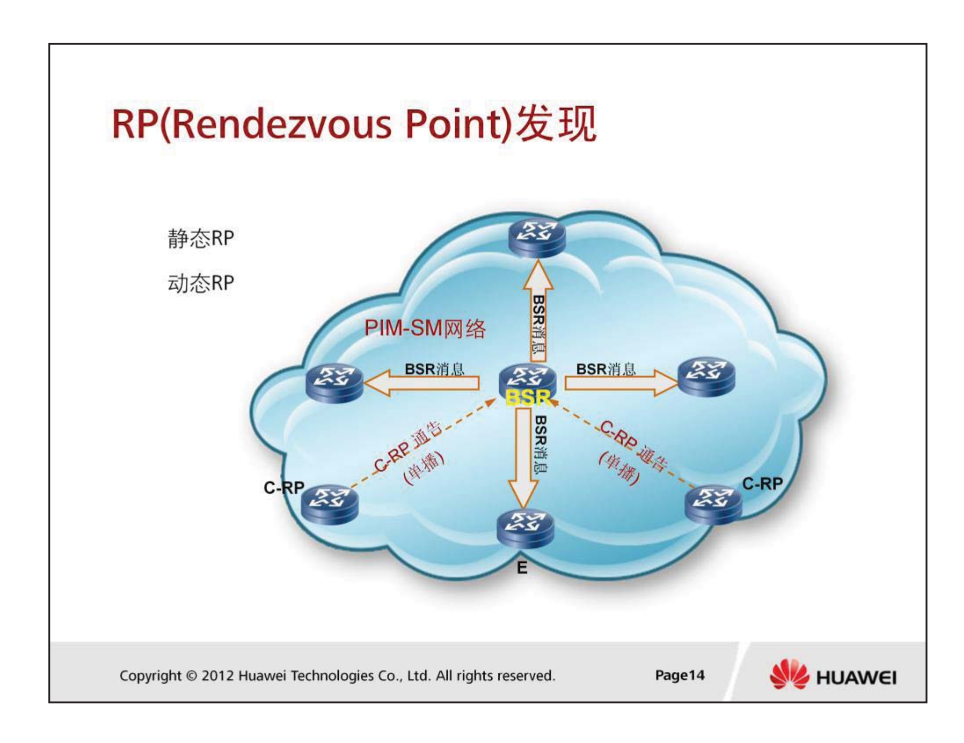
假如优先级相同或网络中至少有一台路由器不支持在Hello报文中携带优先级，则拥有最大IP地址的路由器被选举为DR。

当DR出现故障时，接收Hello消息将会超时，邻居路由器之间会触发新的DR选举过程。

pim timer hello interval，在接口视图下配置发送Hello消息的时间间隔。Hello消息默认周期是30秒。

pim hello-option holdtime interval 命令用于修改Hello消息超时时间值。默认情况超时时间值为105秒。

在共享网络（如Ethernet）有多个路由器时选举出“指定转发器”，PIM-SM中的“Assert”与PIM-DM中的相同，这里不再介绍。



在PIM-SM组播网络里，担当共享树树根的节点称为RP (Rendezvous Point) 。

RP的作用：

1、共享树里所有组播流都通过RP转发到接收者。

2、RP可以负责几个或者所有组播组的转发，网络中可以有一个或多个RP 。用户通过配置命令，可以限制RP只为IP地址在一定范围的组播组服务。一个RP可以同时为多个组播组服务，但一个组播组只能对应一个RP 。所有该组成员和向该组发送组播数据的组播源都向唯一的RP汇聚。

RP的发现：

1、静态RP：在PIM域中的所有PIM路由器上逐一进行配置，静态指定RP 。

static-rp rp-address 指定静态RP的IP地址。

2、动态RP：在PIM域内选择几台PIM路由器，配置成为C-RP(Candidate-RP)，最后从C-RP中竞选产生RP。

使用动态RP，必须同时配置C-BSR（Candidate-BootStrap Router）。由C-BSR竞选产生BSR。

RP是PIM-SM域中的核心路由器，在小型并且简单的网络中，组播信息量少，全网络仅依靠一个RP进行信息转发即可，此时可以在SM域中各

路由器上静态指定RP位置。但是更多的情况下，PIM-SM网络规模都很大，通过RP转发的组播信息量巨大，为了缓解RP的负担同时优化共享树的拓扑结构，不同组播组应该对应不同的RP，此时就需要自举机制来动态选举RP，配置自举路由器BSR(BootStrap Router)。

RP选举原则

如果PIM-SM域中只有一个候选RP（Candidate-RP，C-RP），那么这个节点就是域里的RP。

如果域中存在多个C-RP并都拥有不同的优先级时，则优先级最高（优先级数值越小优先级越高）的将会被选举为域中的RP。

如果域中存在多个C-RP并都拥有相同的优先级时，则依靠Hash算法算出的数值来决定RP，数值最大的成为RP。

- Hash算法参数：
  - 组地址；
  - 掩码长度；
  - C-RP地址。

如果域中存在多个C-RP并都拥有相同的优先级与Hash数值时，则拥有最高IP地址的C-RP为该域的RP。

 Page16



RP利用自举机制来动态选举。

在PIM-SM域中，所有的候选RP(C-RP)参与RP的选举。RP的选举原则如胶片所述。

c-rp interface-type interface-number [group-policy basic-acl-number | priority priority | holdtime hold-interval | advertisement-interval adv-interval] * 命令用于配置C-RP。

interface-type interface-number为C-RP所在接口，该接口必须使能PIM-SM。

group-policy basic-acl-number指定C-RP服务范围为ACL允许的组播组。basic-acl-number表示基本访问控制列表号。缺省情况下，C-RP为所有组播组服务。

priority priority为C-RP的竞选优先级，数值越大，优先级越低。缺省值是0。

在RP选举中，优先级较高的C-RP较优；优先级相同的情况下，执行Hash函数，计算结果较大者获胜；如果Hash函数计算结果也相同，比较IP地址，IP地址较高者较优。

BSR (BootStrap Router)

在PIM-SM网络启动后，负责收集网络内的RP信息，为每个组播组选举出RP，然后将RP集（即组-RP的映射数据库）发布到整个PIM-SM网络的路由器，称之为BSR。

一个PIM-SM域里只有一台BSR，并同时可以存在多台候选BSR（Candidate BootStrap Router，C-BSR）。

 Page17 

BSR自举路由器是PIM-SM网络里的管理核心，负责收集网络中候选RP(C-RP)发来的Advertisement宣告信息。然后将为每个组播组选择部分C-RP信息以组成RP-Set集（即组播组和RP的映射数据库），并发布到整个PIM-SM网络，从而网络内的所有路由器（包括DR）都会知道RP的位置。

C-RP周期性的发送Advertisement宣告消息的时间间隔(advertisement-interval)缺省值是60s。

BSR在holdtime hold-interva(缺省150s)内等待接收C-RP发送的Advertisement宣告消息，超过150s，BSR认为C-RP失效。

一个PIM-SM域内也可以配置多个C-RP，由BSR机制计算出和每个组播组对应的RP。

一个网络（或某管理域）内部只能选举出一个BSR，但可以配置多个候选BSR (Candidate-BSR C-BSR)，当BSR发生故障后，其余C-BSR能够通过自动选举产生新的BSR，从而确保业务免受中断。

BSR的选举

如果域中只有一台C-BSR，该台路由器就是该域里的BSR。

如果域中存在多台C-BSR，则拥有最高优先级的路由器为BSR。

如果域里存在多台拥有相同优先级的C-BSR，则拥有最高IP地址的路由器为BSR。

 Page18

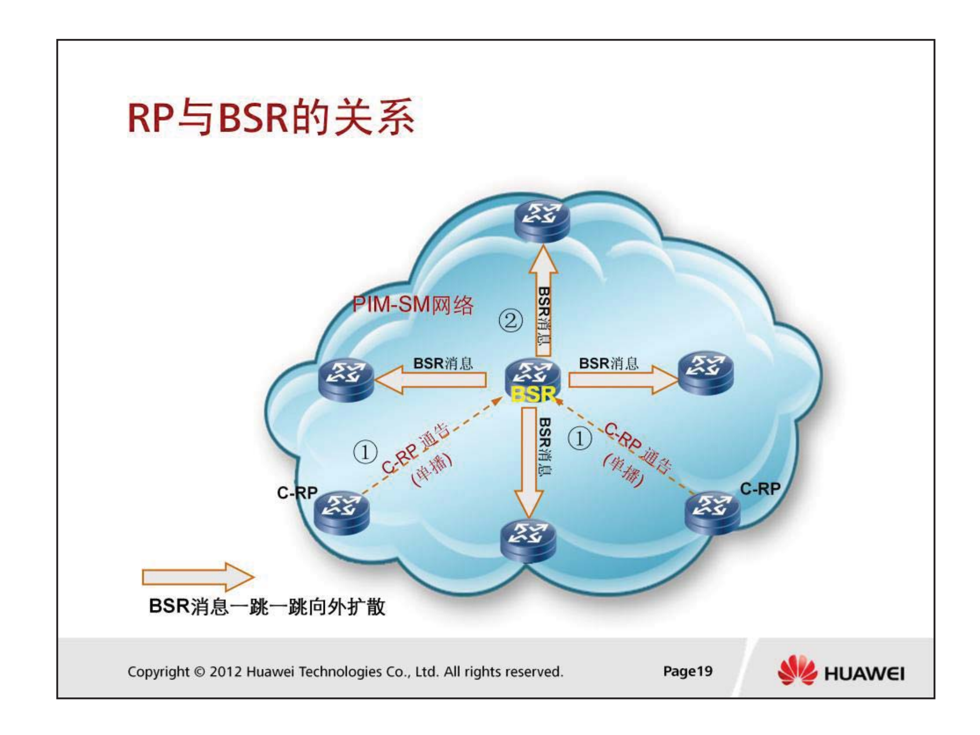


BootStrap router工作的原理和过程：

1、在网络中选择合适的路由器把它配置成候选BSR(C-BSR)，每个C-BSR都有优先级。当它得知自己是C-BSR后，首先启动一个定时器（默认为130秒），监听网络中的 BootStrap Message。BootStrap Message初始时通告发送路由器的优先级、BSR的IP地址。

2、当C-BSR收到一个BootStrap Message后，它会把自己的优先级和报文里的优先级做比较，如果对方的优先级高，它就把自己的定时器重置，继续监听BootStrap Message；如果是自己的高，那么它就发送BootStrap Message声明自己是BSR，如果优先级相等，则比较IP地址，谁的IP地址大谁就是BSR。

BSR消息发送的目的地址是224.0.0.13，所有的PIM路由器都能接收到这个报文，该报文TTL一般被置为1，但每个PIM路由器收到此报文后都是把它以泛洪的方式从自己所有的使能PIM的接口上发送出去，这就能保证网络中的每台PIM设备都能收到BootStrap Message。



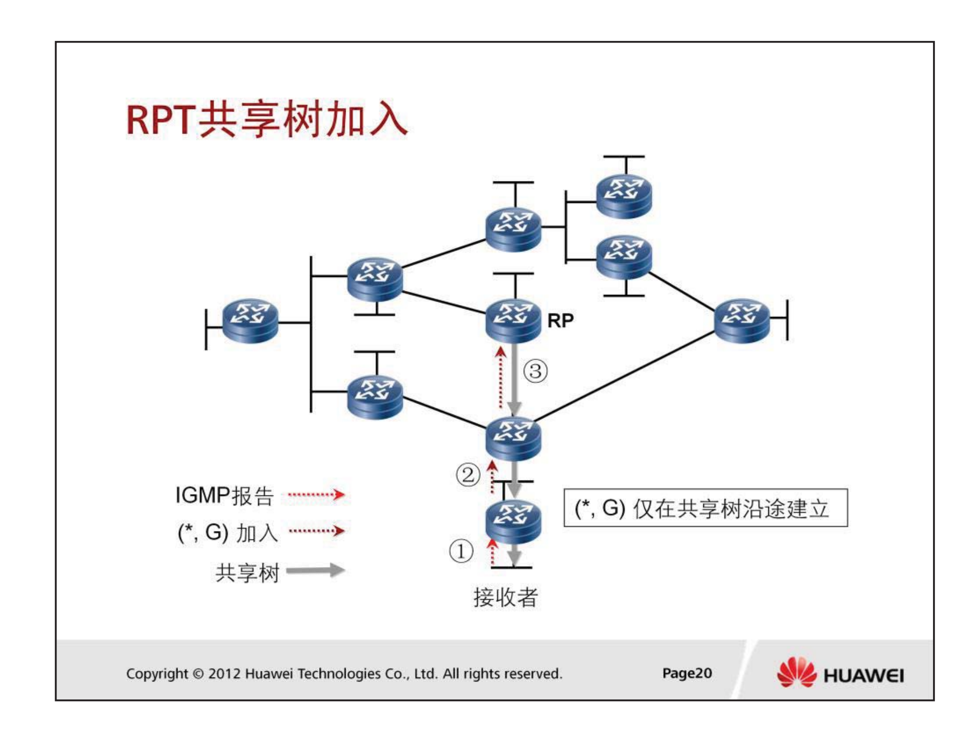
1、候选RP(C-RP)周期性将声明发送到BSR。

当C-RP收到BootStrap Message后，它可以从该message中得知网络中谁是BSR，然后C-RP通过Candidate-RP-Advertisement Message把自己所能服务的组单播给BSR。每个C-RP都这么做的话，BSR就收集到了网络中所有C-RP的信息，并把这些信息整理成一个集RP-Set。C-RP每60秒周期性的单播发送通告。

2、BSR通过BootStrap Message周期性地向所有PIM路由器(224.0.0.13)发送BSR消息(每60秒)，BSR消息包含整个RP-set和 BSR地址，消息一跳一跳地自BSR向整个网络泛洪（flood）。

3、所有的路由器使用收到的RP集来确定RP。所有路由器都使用相同的RP选择算法，所以选择的RP也是一致的。

注意：如果RP不是手工指定，而是通过选举从C-RP中产生，则每台路由器需要配置包括C-RP地址、优先级和它所能服务的组。



当接收者主机加入一个组播组G时，通过IGMP报文知会与该主机直接相连的叶子路由器，叶子路由器掌握组播组G的接收者信息，然后朝着RP方向往上游节点发送加入组播组的Join消息。

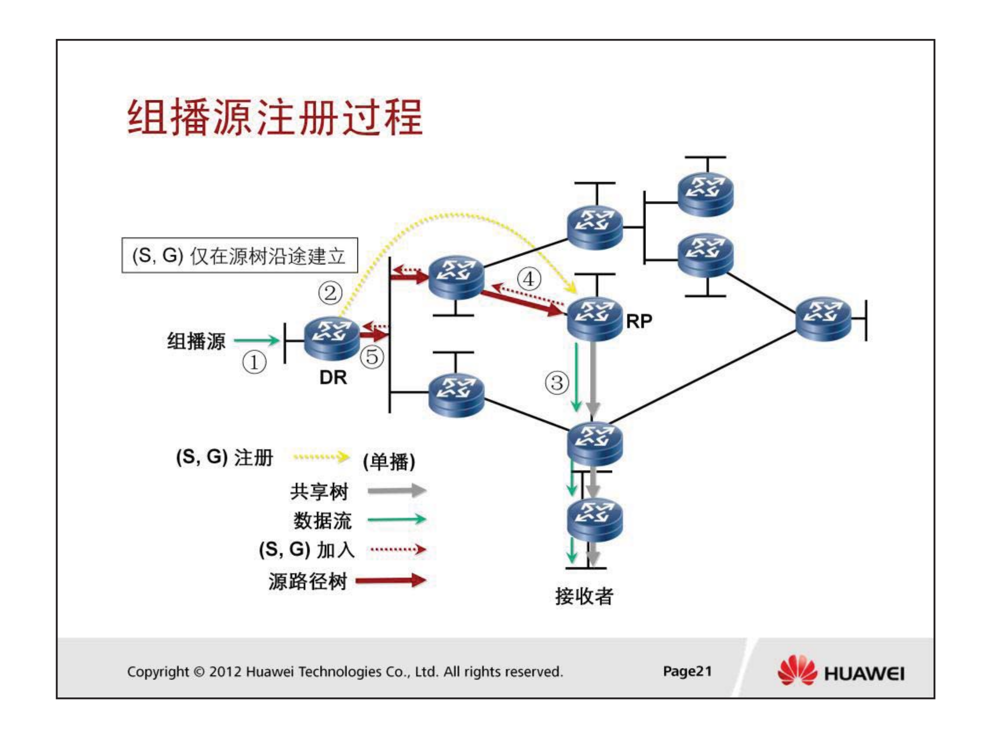
从叶子路由器到RP之间途经的每个路由器都会在转发表中生成(*，G)表项，这些沿途经过的路由器就形成了RP共享树(RPT)的一个分支。其中(*，G)表示从任意源来的信息去往组播组G。

RPT共享树以RP为根，以接收者为叶子。

当从组播源Source来的发往组播组G的报文流经RP时，报文就会沿着已经建立好的RPT共享树路径到达叶子路由器，进而到达接收者主机。

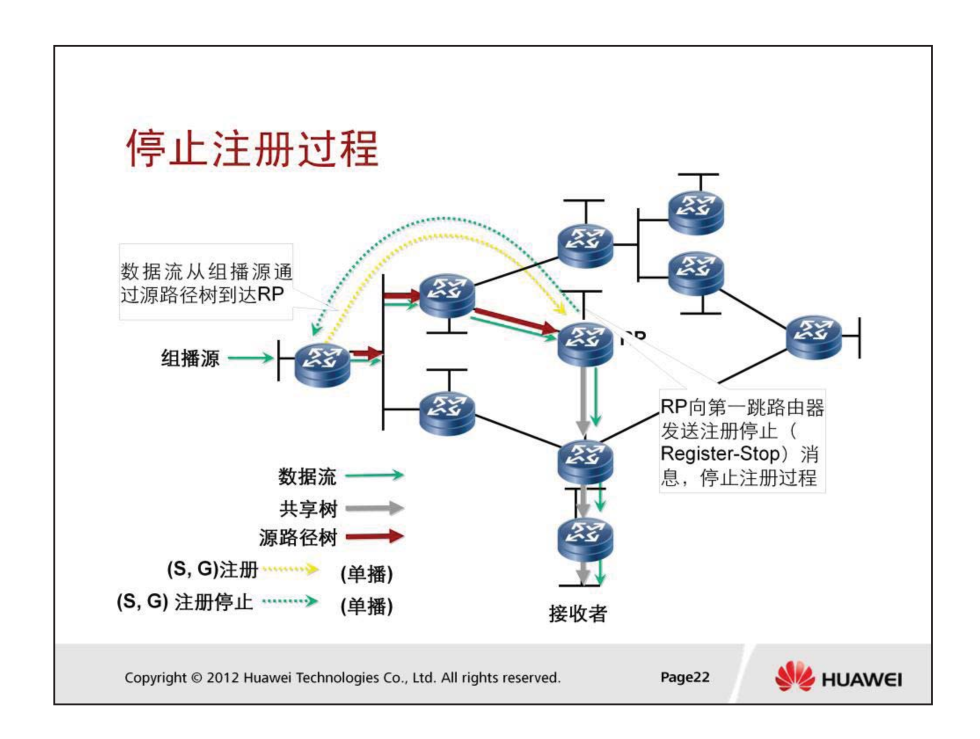
当某接收者对组播信息不再感兴趣时，离该接收者最近的组播路由器会逆着RPT树朝RP方向逐跳发送Prune剪枝消息。第一个上游路由器接收到该剪枝消息，在接口列表中删除连接此下游路由器的接口，并检查自己是否拥有感兴趣的接收者，如果没有则继续向上游转发该剪枝消息。

这一过程同PIM-DM的剪枝相同。



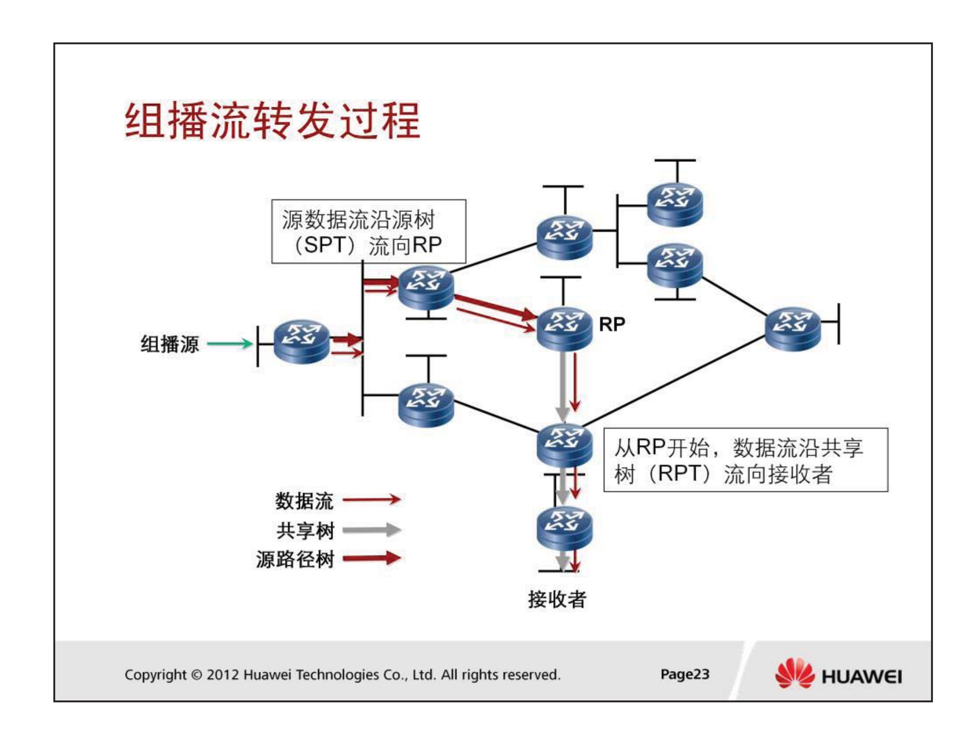
为了向RP通知组播源的存在，当组播源向组播组G发送了一个组播报文时，与组播源直接相连的路由器接收到该组播报文后，就将该报文封装成Register注册报文，并单播发送给对应的RP。

当RP接收到来自组播源的注册消息后，一方面解封装注册消息并将组播信息沿着RPT树转发到接收者，另一方面朝组播源方向逐跳发送(S，G)加入消息，从而让RP和组播源之间的所有路由器上都生成了(S，G)表项，这些沿途经过的路由器就形成了SPT树的一个分支。SPT源树以组播源为根，以RP为目的地。

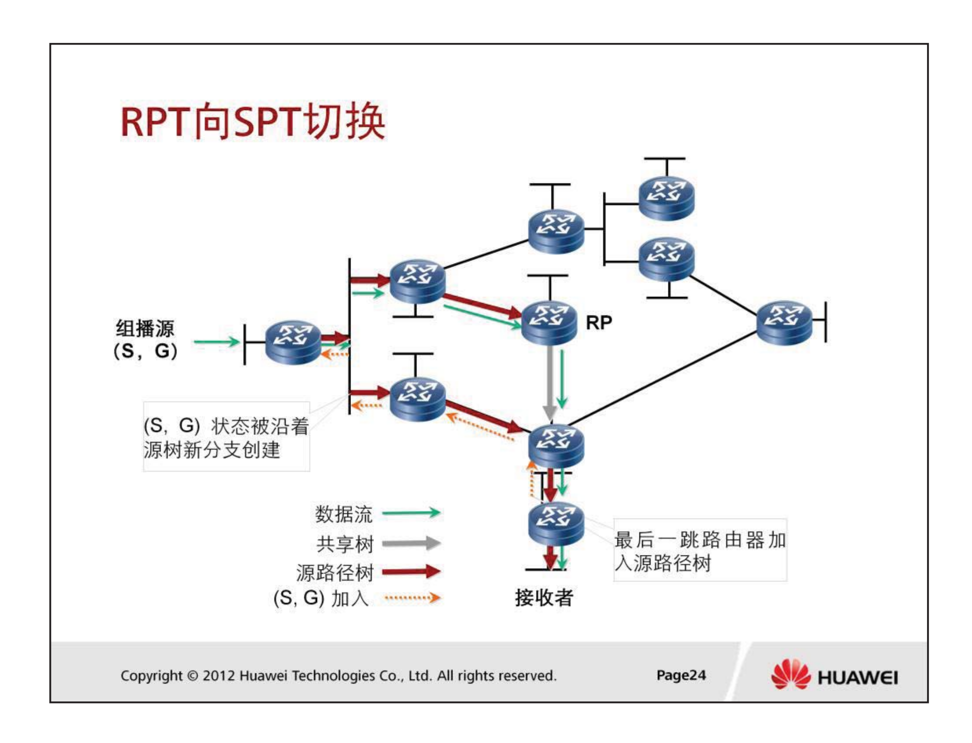


组播源发出的组播信息沿着已经建立好的SPT树到达RP，然后由RP将信息沿着RPT共享树进行转发。

当RP收到沿着SPT树转发的组播流量后，向与组播源直连的路由器单播发送注册停止报文。组播源注册过程结束。



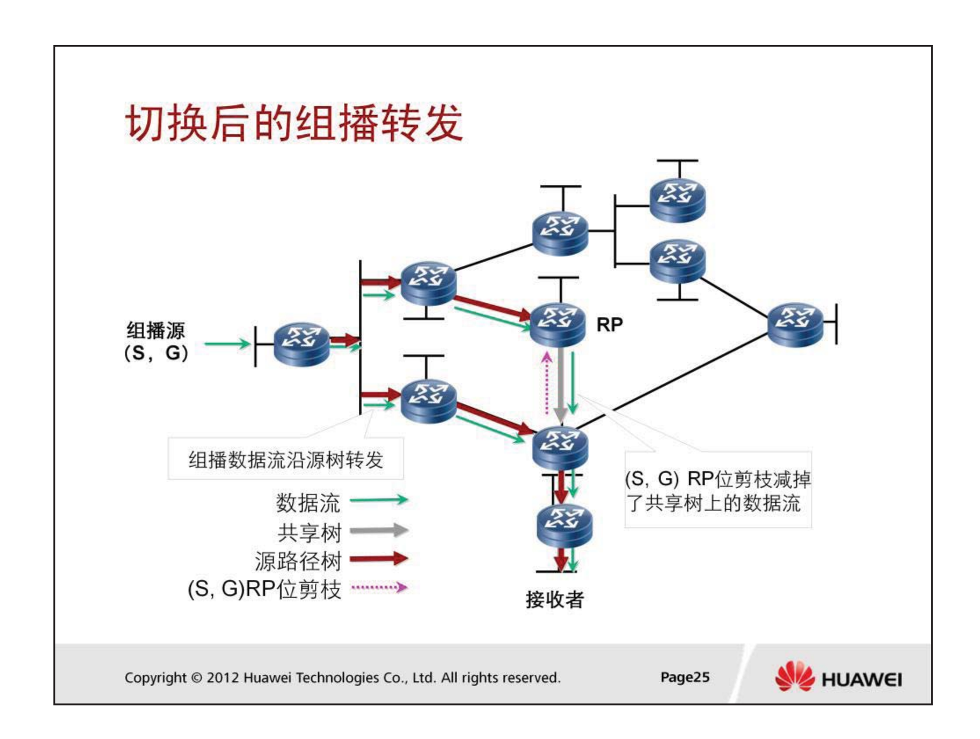
当组播源成功注册到RP后，组播报文从组播源经SPT树到RP，再由RP经RPT树向接收者方向转发。



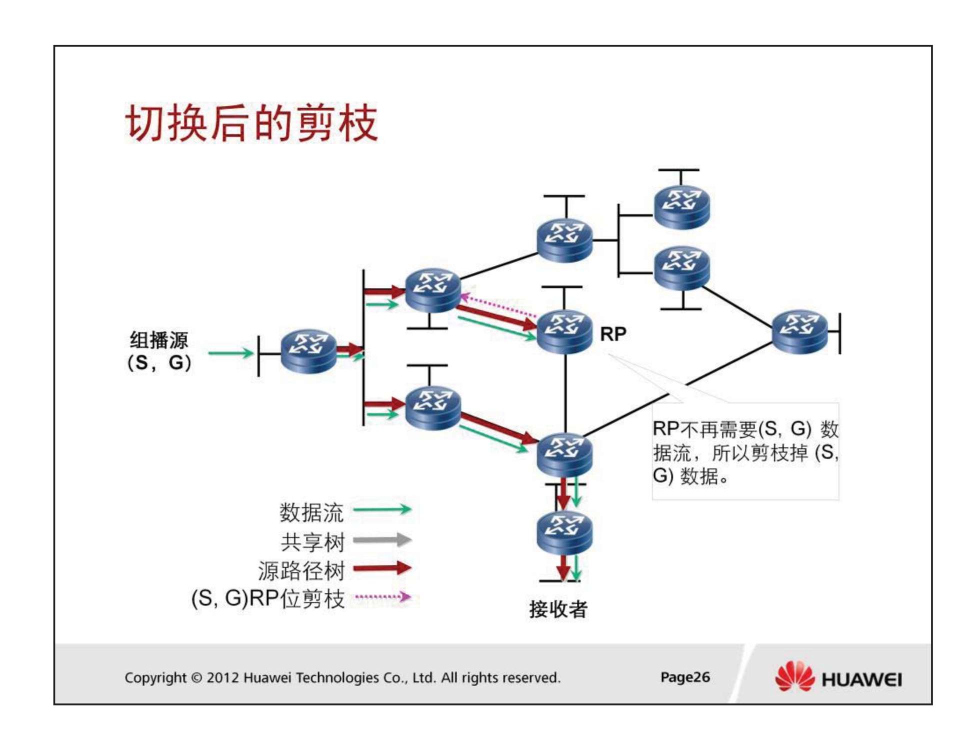
针对特定的源，PIM-SM通过指定一个利用带宽的SPT阈值可以实现将最后一跳路由器(即离接收者最近的DR)从RPT切换到SPT。当最后一跳路由器发现从RP发往组播组G的组播报文速率超过了该阈值时，就向单播路由表中到组播源S的下一跳路由器发送（S，G）加入消息，Join加入消息经过一个个路由器后到达第一跳路由器（即离组播源最近的DR），沿途经过的所有路由器都拥有了（S，G）表项，从而建立了SPT树分支。

当信息吞吐率超过预定的值时，PIM-SM就会从共享树切换到组播源路径树。

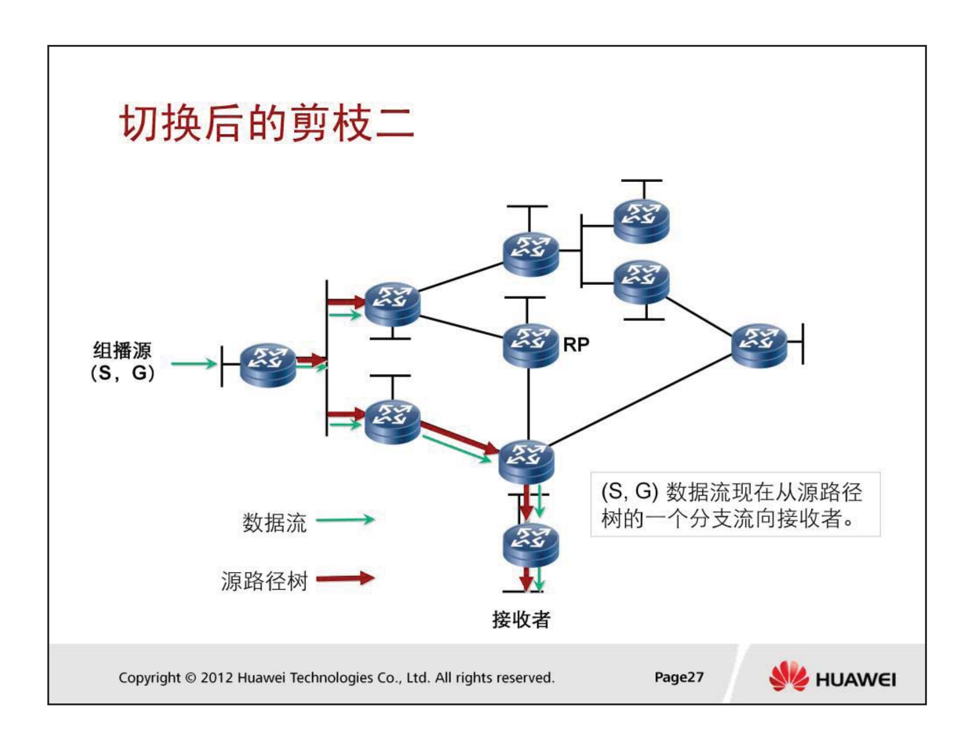
在VRP中，缺省情况下连接接收者的路由器在探测到组播源之后（即接收到第一个数据报文），便立即加入最短路径树（源树），即从RPT向SPT切换。



切换到SPT树后，组播信息将直接从组播源S发送到接收者。通过RPT树到SPT树的切换，PIM-SM能够以比PIM-DM更经济的方式建立SPT转发树。



最后一跳路由器向RP逐跳发送包含RP位的Prune剪枝消息，RP收到消息后会向组播源反向转发Prune剪枝消息，从而最终实现组播信息流从RPT树切换到SPT树。



切换后从组播源到接收者之间建立了SPT。

PIM-SM工作机制

PIM-SM的工作过程主要有：

- 邻居发现
- DR选举
- RP发现
- 加入
- 剪枝
- 注册
- SPT切换

 Page28 

邻居发现：在PIM-SM网络中，组播路由器使用Hello消息来发现邻居，并维护邻居关系，协商协议参数。通过比较Hello消息上携带的优先级和IP地址，各路由器为多路由器网段选举指定路由器DR，充当IGMPv1的查询器。

DR选举：为与组播源或组播接收者之间的共享网络（如Ethernet）选举DR（Designated Router）。

RP发现：通过手工指定或是通过BSR自举消息选举产生。

加入（Join）：当接收者加入一个组播组G时，通过IGMP报文知会与该主机直接相连的叶子路由器，叶子路由器朝着RP方向往上游节点发送加入组播组的Join消息。

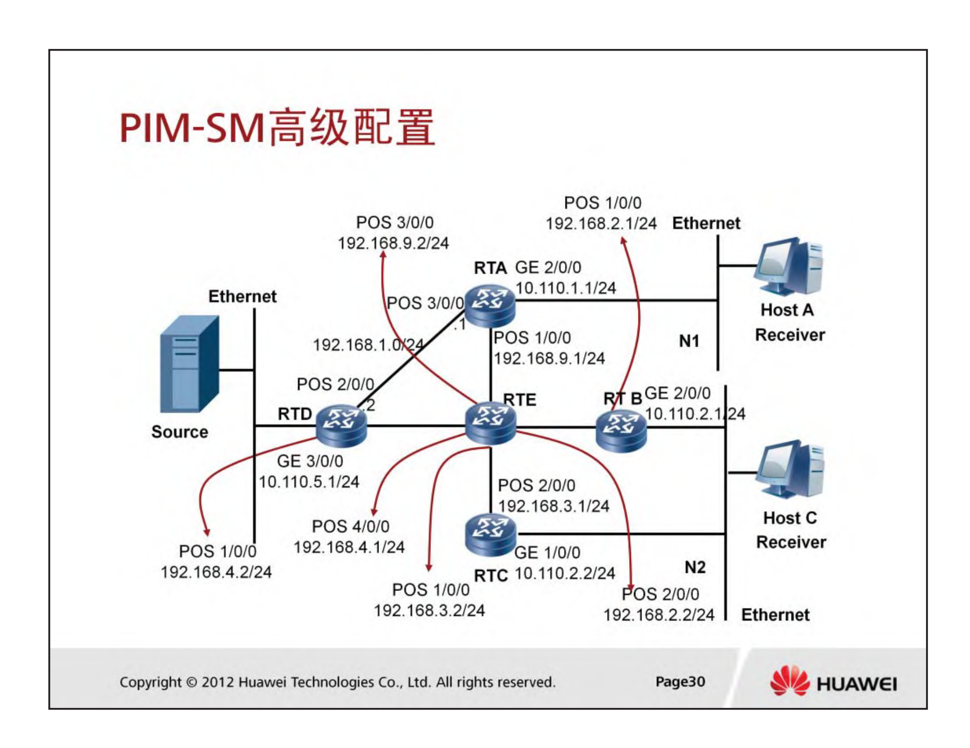
剪枝（Prune）：剪枝过程最先由叶子路由器发起。下游组播组成员全部离开，则向上游节点发Prune剪枝消息。通知上游节点不用再转发数据到该分支。

注册（Register）：向RP通知组播源S的存在。

SPT切换：PIM-SM通过指定一个利用带宽的SPT阈值可以实现将最后一跳路由器(即接收者侧DR)从RPT切换到SPT。



本章举例说明PIM-SM的高级配置，即通过BSR选举产生RP的配置。



RTE的POS3/0/0接口作为此PIM-SM网络的C-BSR和C-RP，运行IGMPv2。

配置思路：

1、配置各路由器的接口IP地址和OSPF路由协议，保证任意网段的路由可达。

2、各路由器使能组播路由功能。

3、各路由器接口上使能PIM-SM，主机侧接口上使能IGMP功能。

4、配置RTE的POS3/0/0接口为C-BSR和C-RP。

PIM-SM基本配置

```
[RTA] multicast routing-enable
[RTA] interface gigabitethernet 2/0/0
[RTA-GigabitEthernet2/0/0]igmp enable
[RTA-GigabitEthernet2/0/0]igmp version 2
[RTA-GigabitEthernet2/0/0]pim sm
[RTA-GigabitEthernet2/0/0]quit
[RTA] interface pos 3/0/0
[RTA-Pos3/0/0]pim sm
[RTA-Pos3/0/0]quit
```


RTB、RTC、RTD和RTE上的配置过程与Router A上的配置相似。

C-RP和C-BSR配置

```
[RTE]acl number 2005
[RTE-acl-basic-2005]rule permit source 225.1.1.0 0.0.0.255
[RTE-acl-basic-2005]quit
[RTE]pim
[RTE-pim]c-bsr pos 3/0/0
[RTE-pim]c-rp pos 3/0/0 group-policy 2005
```

 Page32 

[RTE]pim 全部使用PIM协议 。

[RTE-pim]c-bsr pos 3/0/0 配置C-BSR所在的接口为RTE的Pos3/0/0

[RTE-pim]c-rp pos 3/0/0 group-policy 2005配置C-RP所在的接口为RTE的Pos3/0/0，并指定C-RP的服务范围为225.1.1.0网段的组播组。

缺省情况下，C-RP为所有组播组服务。如果想指定C-RP的服务范围，可以通过配置策略来实现，具体是通过配置ACL列表来设置允许服务的组播组。

[RTE]acl number 2005 创建ACL列表

[RTE-acl-basic-2005]rule permit source 225.1.1.0 0.0.0.255 配置允许的地址范围。

c-rp interface-type interface-number [group-policy basic-acl-number | priority priority | holdtime hold-interval | advertisement-interval adv-interval] *

interface-type interface-number 配置C-RP所在接口，该接口必须使能PIM-SM。

参数说明:

group-policy basic-acl-number 指定C-RP服务范围为ACL允许的组播组。basic-acl-number表示基本访问控制列表号。

priority priority 为C-RP的竞选优先级，数值越大，优先级越低。缺省值是1。

undo c-rp { interface-type interface-number | all }取消c-rp配置。

c-bsr interface-type interface-number hash-mask-len [priority] 配置C-BSR所在的接口，掩码长度和优先级。

参数说明:

interface-type interface-number：指定候选BSR所在的接口，该接口一定要启用PIM-SM，配置才能生效。

hash-mask-len：指定掩码长度，该掩码先和组播地址进行“与”操作，然后再进行查找RP的操作。取值范围为0～32。

priority：该候选BSR的优先级。优先级的数值越高候选自举路由器的优先级越高。取值范围为0～255。缺省优先级为0。

undo c-bsr用来取消候选BSP配置。

PIM-SM配置验证

查看路由器接口上PIM的配置和运行情况

```
[RTA]display pim interface
Interface     NbrCnt HelloInt   DR-Pri    DR-Address
GE2/0/0         0       30        1       10.110.1.1 (local)
Pos3/0/0        1       30        1       192.168.1.2
Pos1/0/0        1       30        1       192.168.9.2
```




命令display pim interface用来显示所有的PIM接口信息。使能了PIM-SM协议的接口，都可以被查看到。

PIM-SM配置验证

使用display pim bsr-info命令可以查看路由器上BSR选举的信息

```
[RTA] display pim bsr-info
    Elected BSR Address: 192.168.9.2
          Priority: 0
          Hash mask length: 30
          State: Accept Preferred
          Scope: Not scoped
          Uptime: 01:40:40
          Expires: 00:01:42
```

 Page35 HUAWEI

命令display pim bsr-info用来显示自举路由器（BSR）信息。

包括BSR的地址，哈希掩码长度，优先级。



命令display pim rp-info用来显示所有的组播组对应的RP信息，包括Auto-RP机制发现的RP、BSR机制发现的RP和静态RP的信息。

PIM-SM配置验证

使用display pim routing-table命令查看路由器PIM协议组播路由表

```
[RouterA] display pim routing-table
(*, 225.1.1.1), RP: 192.168.9.2
Protocol: pim-sm, Flag: WC
UpTime: 00:13:46
Upstream interface: Pos1/0/0, RPF neighbor: 192.168.9.2
Downstream interface list:
GigabitEthernet2/0/0, Protocol: static, UpTime: 00:13:46,
  Expires:-
(10.110.5.100, 225.1.1.1), RP: 192.168.9.2
Protocol: pim-sm, Flag: SPT LOC
UpTime: 00:00:42
Upstream interface: pos3/0/0, RPF neighbor: 192.168.1.2
Downstream interface list: GigabitEthernet2/0/0
```

 Page37 HUAWEI

display pim routing-table用来显示所有实例的PIM协议组播路由表的内容。

显示了（*，G）表项和（S，G）表项中分别对应的上行接口，RFP邻居和下行接口信息，RP信息。

（*，G）和（S，G）表项的建立说明了在本路由器上组播数据可以正确转发。



Page38



1、PIM-SM的工作原理?

PIM-SM协议假设：当组播源开始发送组播数据时，域内所有的网络节点都不需要接收数据。实现组播转发的核心任务是构造并维护一棵单向共享树。共享树选择PIM中某一路由器作为RP。组播数据通过RP沿着共享树向接收者转发。在接收侧，连接信息接收者的路由器向该组播组对应的RP发送组加入消息，加入消息经过一个个路由器后到达根部（即RP汇聚点），所经过的路径就变成了此共享树RPT的分支。发送端如果想要往某组播组发送数据，首先由第一跳路由器向RP汇聚点进行注册，注册消息到达RP后触发源树建立。之后组播源把数据发向RP汇聚点，当数据到达了RP汇聚点后，组播数据包被复制并沿着RPT树传给接收者。复制仅仅发生在分发树的分支处，这个过程能自动重复直到数据包最终到达接收者。

2、如何建立RPT共享树?

当接收者主机加入一个组播组G时，通过IGMP报文知会与该主机直接相连的叶子路由器，叶子路由器然后朝着RP方向往上游节点发送加入组播组的Join消息。从叶子路由器到RP之间途经的每个路由器都会在转发表中生成（*，G）表项，沿途经过的路由器就形成了RP共享树（RPT）的一个分支。

3、组播源如何注册?

当组播源S向组播组G发送了一个组播报文时，与组播源S直接相连的路由器接收到该组播报文后，就将该报文封装成Register注册报文，并单播发送给对应的RP。 当RP接收到来自组播源S的注册消息后，一方面解封装注册消息并将组播信息沿着RPT树转发到接收者，另一方面朝组播源S逐跳发送（S，G）加入消息，从而让RP和组播源S之间的所有路由器上都生成了（S，G）表项，沿途经过的路由器就形成了SPT树的一个分支。SPT源树以组播源S为根，以RP为目的地。


谢谢
www.huawei.com

在线学习资料支持

您可以在华为企业业务网站获得E-Learning课程、培训教材 、产品资料、软件工具、技术案例等：

1、 E-Learning课程：登录*华为在线学习网站*，进入“*华为培训/在线学习*”栏目

免费E-Learning课：对网站所有用户免费开放

职业认证E-Learning课：通过任何一项职业认证即可学习所有职业认证培训E-Learning课程

渠道赋能E-Learning课：对华为企业业务合作伙伴免费开放

2、培训教材：登录*华为在线学习网站*，进入“*华为培训/面授培训*”， 在具体课程页面即可下载教材。

华为职业认证培训教材、华为产品技术培训教材。无需注册即可下载

3、华为在线公开课(LVC)： http://support.huawei.com/ecommunity/bbs/10154479.html

企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师公开授课

4、产品资料下载： http://support.huawei.com/enterprise/#tabname=productsupport

5、软件工具下载：http://support.huawei.com/enterprise/#tabname=softwaredownload

更多内容请访问：

- http://learning.huawei.com/cn
- http://support.huawei.com/enterprise/
- http://support.huawei.com/ecommunity/

HUAWEI